Population genetic analysis of two species of non-indigenous riparian weeds in northeast England in the context of their spatial ecology

Walker, Nathalie Fleur

Population Genetic Analysis of Two Species of Non-indigenous Riparian Weeds in Northeast England in the Context of Their Spatial Ecology

by

Nathalie Fleur Walker

Department of Biological Sciences,

University of Durham.

2000

This thesis is submitted in

candidature for the degree of

Doctor of Philosophy

Declaration

The material contained within this thesis has not previously been submitted for a degree at the University of Durham or any other university. The research reported within this thesis has been conducted by the author unless indicated otherwise.

Abstract

The population genetic structure of two species of invasive non-indigenous riparian weeds in the Northeast of England was investigated using microsatellite markers. *Heracleum mantegazzianum* and *Impatiens glandulifera* were introduced into the UK from Asia. The first records of the species in the Tees, Tyne and Wear catchment areas were in 1944 and 1892 respectively. Both species have spread rapidly, and are present over a wide area of the catchments. The pattern of genetic variation was investigated in order to determine the importance of anthropogenic introduction, and life-history and dispersal strategies to the distribution of the species. Twelve populations of each species were sampled from the Tees, Tyne and Wear catchments as well as an independent population for comparison. Genomic libraries were constructed and screened for dinucleotide repeat microsatellite loci. Four polymorphic loci of *H. mantegazzianum* and three of *I. glandulfera* were identified, and each species was also screened for variation using one universal chloroplast microsatellite locus. A large amount of variation was found in both species as the loci of *H. mantegazzianum* had between nine and twenty alleles and those of *I. glandulifera* between eight and sixteen alleles. Results revealed greater overall variation between populations from different catchments than those in the same catchment. Within a catchment, there was evidence of isolation by distance for both species in one out of two catchments examined. Populations of *I. glandulifera* showed greater temporal variation and there was more variation both overall and within a catchment in this species. This is likely to be due to the larger number of individuals present, and the wider distribution of this species. Low levels of chloroplast variation were found in both species. This may reflect a lack of variation in the material introduced into the UK.

Acknowledgements

I would like to firstly thank my supervisors, Rus Hoelzel and Phil Hulme, for their constant support, and for always being available whenever their help was requested.

Thanks are owed to Yvonne Collingham for providing me with first records of my study species, and to Vicky Kelly and Gillian Storey for all their work in the DNA sequencing lab. I am very grateful to Julia Bartley, who kindly taught me how to process GeneScan™ gels, saving me weeks of trial and error. I would like to thank Dan Engelhaupt, Anna Fabiani, Stefania Gaspari, Diane Hart, Ada Natoli and Ana Topf, for their help in the molecular ecology lab, and Mark Skipsey, David Dixon and Ian Cummings for their resourcefulness and remarkable generosity with their time and lab equipment. The Environment Agency was a great help in providing River Corridor Survey data.

Steve Willis warrants many thanks for his help with ecological data, finding roots in the middle of winter, tips on how to avoid *Heracleum* rashes and for being such a good friend. Steve Woodhouse and Prash Sinnathamby were a great support, and writing this thesis would have been much less enjoyable without the entertainment of Geoff Busswell, Hendrick Dey, Sarah Fahle, Rich Hutchinson, Stinson Lindenzweig, Jim O'Donnell, Alex Porter, Rosie Sarrington, Chris Stokoe and Ed Sumner.

I am very grateful to my sister and grandmother, particularly for being so helpful to me whilst convalescing from a tonsil operation, and to Richard Sherrington for his very generous IT support.

Finally, I would like to thank my parents, to whom this thesis is dedicated.

Table of Contents

Chapter Five– Results II: *Impatiens glandulifera*

Chapter Six - Discussion

List of Figures

List of Tables

Chapter One

Introduction

1.1 Invasive Species

Invasions by non-indigenous species have made a large impact on natural communities, particularly on islands and can affect the structure and lower the biodiversity of ecosystems (Pysek & Pysek 1995). Tahiti has been greatly affected by introduced plants, with nearly half the endemic species at risk of extinction (Meyer & Florence 1998). *Miconia calvescens,* which is native to Latin America, was introduced as an ornamental plant but has spread over two-thirds of Tahiti and the species is a threat to other Pacific islands.

There are many more examples of non-indigenous species that have had serious effects on native species and ecosystems across wide areas. Eight species of *Tamarix* (salt cedar) were first introduced to North America in the 1800s from the Mediterranean and they are now a major threat to riparian woodland because of their high reproductive ability and their affect on soils and hydrology (Di Tomaso 1998). It is able to increase the salt concentration in soils around it to levels that native species are unable to tolerate. *Sapium sebiferum* (Chinese tallow) was introduced into the United States from Asia and is now widespread across the Southeastern states (Barrilleaux & Grace 2000). Its shed leaves contain toxins which affect soil chemistry and enable it to displace native vegetation. The zebra mussel (*Dreissena polymorpha*) has spread throughout most of eastern North America in just fifteen years and the costs of its damage runs into billions of dollars. The cinnamon fungus, *Phytophthora cinnamomi* threatens to severely reduce the biodiversity of the Australian bush (Kirkpatrick 1994). Most garden plants are resistant to the fungus, which is thought to have been introduced into the wild by the dumping of garden waste into the bush.

There are over 900 non-indigenous plant species that have either been naturalised (permanently established) or persist by frequent reintroduction into the British Isles and they comprise over 40 % of the total number of plant species (Lovei 1997). A large number of non-indigenous species are introduced into Britain as crops or garden plants. Many are accidentally introduced as seeds mixed in with grain, or in ships'

ballasts. The vast majority of introduced species fail to become established because the habitat or climate is unsuitable, they are killed off by pests and diseases, or they cannot compete successfully with the native species. Even if a species can survive in a new area, small founder populations are prone to decline due to demographic stochasticity (Crawley 1989a). The proportion of species introduced that are likely to become a serious threat to native habitats can be estimated from the tens rule (Williamson & Brown 1986). This states that 10% of imported plants escape into the wild, 10% of these become established and 10% of those established become serious pests. Species that become established in new areas are therefore of considerable interest.

The absence of specialised herbivores, seed predators, parasites and competitors in invaded areas can lead to the very rapid spread of non-indigenous species. These species can cause considerable damage to natural communities and in some cases great efforts are made to control them. *Salvinia molesta,* and aquatic fern native to South America, invaded large areas of Africa and Southeast Asia. Its spread was aided by the lack of specialised herbivores in invaded areas (Williamson 1986). When a native beetle, *Cyrtobagous salviniae,* was used as a biological control agent, populations in Australia, Namibia, India and Papua New Guinea were greatly reduced.

Holdgate (1986) generalised invaders as being r-strategists that produce large numbers of seeds, have large populations, disperse over long distances, persist in a wide variety of habitats, tolerate environmental fluctuations, and are aided by disturbance, often by humans. These factors are often interrelated and the most important attributes vary in different cases. Dispersal ability may be a limiting factor for the persistence of a species that requires specialised habitats and has small populations because other suitable habitats may be widely scattered (Skogland 1990). However, for species with more generalist habitat requirements, that have high population densities, dispersal ability is less important for persistence as seedlings can establish in areas near to the main stand. Once established, the impact of an invader also depends upon its intrinsic rate of increase, the rate of arrival of new immigrants, the level of interspecific competition, and the effect of pests, diseases and mutualists.

The impact of invasions depends just as much on the make up and conditions of sites reached by alien species as it does on the attributes of the invaders themselves (Pysek & Pysek 1995). Habitats most likely to be invaded are disturbed areas, or man-made sites with few native species, little plant cover, low nutrient levels or unstable soils and small seed banks (Crawley 1986). Riparian areas are amongst the most

common places for invasions to occur as they are often disturbed by flooding, banks may be unstable and they are susceptible to invaders as seeds floating down rivers become deposited on banks.

The study of invasions can give an insight into the processes that shape the distribution and abundance of species. It can be difficult to understand the reasons for a species' distribution in space and time, because occurrences that have influenced this either take place too slowly to measure or have taken place in the past and often cannot be accounted for. By studying the spread of an invader that has yet to reach its potential distribution, these problems can be overcome and factors such as interspecific competition and unusual disturbances can be taken into account (Mack 1985).

Following the establishment of a species in a new area, there is a slow initial spread, known as the 'lag' phase, when the species occurs only in a few areas. This is followed by rapid spread in a phase of expansion that continues until the limits of distribution are reached. This then leads to the final phase in which there is little spread. The spread usually starts from a number of foci which either come about from independent introductions or by a founder population establishing a species in an area on several occasions (Moody & Mack 1988). Foci that are originally small can, if suitably situated, be of great importance in the subsequent rate of expansion of a species. The study of genetic variation in a species may lead to the identification of such foci.

The pattern of spread of an invader is affected by a number of different factors, which make predictions difficult. These factors include the size of the colonising population, the number of initial foci, the timing of an invasion, the suitability of the surrounding habitat, the heterogeneity of the area and the number and size of corridors and barriers to spread. The life history strategy of an invader is also important. Species that reproduce purely by vegetative means may face problems due to low genetic variation (which can lead to an inability to survive environmental changes) and their inability to produce seeds (that may disperse or form a seedbank). Dioecious species require the presence of plants of both sexes among founder populations. Perennials that require a long time to reach reproductive age may find it particularly difficult to persist. Another constraint on the ability to predict the spatial pattern of invasions is human-mediated long distance dispersal, but perhaps the major difficulty is the important role played by chance. This can be divided into environmental and demographic stochasticity (Crawley 1989b). Extreme weather events can greatly

affect invaders, which may have never before been exposed to such conditions, eg. *Ailanthus altissima* in Provence. Native species are more likely to survive environmental fluctuations given their previous exposure to such events and their consequent adaptation. The intrinsic rate of increase for a species is determined by fecundity and mortality, both of which can be affected by chance and timing, such as the age of immigrants, and the availability of resources at the time of invasion (Freckleton and Watkinson 1998).

Two species of invasive riparian plants native to Asia, *Heracleum mantegazzianum* (Sommier & Levier) and *Impatiens glandulifera* (Royle), have widespread distributions in the UK and were introduced into Northeast England in the last century.

1.2 *Heracleum mantegazzianum*

Heracleum mantegazzianum is a diploid monocarpic perennial native to southwest Asia that has spread west throughout northern Europe and to North America. It has been recorded in the British Isles from Cornwall to the Highlands of Scotland and in Ireland (Tiley *et al.* 1996). The earliest record of its presence in the wild in Britain was in 1893. It spread slowly at first, but its distribution has increased by more than 40 times in the last 50 years. Its first record in the Tees, Tyne and Wear catchment area was in 1928 near Hexham on the Tyne (Biological Records Centre, Monkswood). The first record of *H. mantegazzianum* in the Wear catchment was in 1954 and the species was first recorded in the Tees catchment in 1944 (Biological Records Centre, Monkswood). *Heracleum mantegazzianum* is the largest forb in Europe and its size and high growth rate makes it very competitive, forming dominant stands where it occurs, shading out all competitors. It has spread at a rapid rate, having escaped from gardens with the aid of humans through the collection of seeds. It is most prevalent in Britain on riverbanks, the second most common site being ruderal areas. It is particularly successful in open areas because the resident species are likely to be light-dependent and are easily outcompeted once they become shaded out.

Within a stand, there is much competition for light and only the tallest plants flower (Tiley *et al.* 1996). Plants take two to five years to flower, with most flowering in the third year. The flowers are arranged in compound umbels. Umbels occur in

groups, with a terminal umbel being surrounded by up to eight satellite umbels. There may be a number of terminal umbels on one plant. Reproduction is entirely sexual and within an umbel, the stigma is not ready to receive pollen until after its own pollen has been shed. However, there can be an overlap in the timing of producing and accepting pollen between terminal and satellite umbels and between terminal umbels. Therefore, *H. mantegazzianum* can self-pollinate. It is pollinated by a variety of insects, mainly from the families of Diptera and Hymenoptera. Following pollination, the fruit divides into two winged mericarps. The mean dry weight of mericarps measured in a range of studies vary from 5.7 mg to 16.5 mg (Tiley *et al.* 1996). The seed size and weight depends on the condition of the plant and the location of the umbel from which they arose, with the largest seeds stemming from the terminal umbel. Estimates of the number of seeds produced by each plant range from 1,500 to over 100,000. Following seed set, the whole plant dies.

Seeds are dispersed short distances by wind and longer distances by water. Most seeds land within 10 m of a stand, but seeds reaching rivers can float (possibly for days) downstream, before landing on areas of deposition, where they may germinate as they can tolerate waterlogged soils. The maximum dispersal distance for a single propagule has been estimated as 10 km (Wadsworth *et al.* 2000). However, because of the nature of currents and eddies, the majority of seeds reaching rivers are unlikely to have travelled further than 50 m (Tiley *et al.* 1996).

Heracleum mantegazzianum can cause contact dermatitis when touched by human skin and this problem has fuelled efforts for its control in several parts of Europe (Pysek & Pysek 1995).

1.2.1 Hybridisation of *H. mantegazzianum* with *H. sphondylium*

Heracleum mantegazzianum can hybridise with *H. sphondylium* although no hybridisation has been shown to occur outside the British Isles (McClintock 1975). Hybrids are intermediate in character between the two species (McClintock 1975; Arora *et al.* 1982; Tiley *et al.* 1996). They have low pollen fertility and are expected to be present at a frequency of less than 0.1% where the two species occur together (Grace & Nelson 1981). Weimark *et al.* (1979) found that hybrids differed morphologically from *H. mantegazzianum* especially in height, number of rays in the terminal umbel and in basal stem diameter. The two species can be pollinated by the same insect species, but selective foraging of pollinators leads to very few insects

carrying pollen from both species (Grace & Nelson 1981). This may be due to insects exhibiting 'search-image' behaviour and would limit the likelihood of cross-pollination, and so could be partly responsible for the low numbers of hybrids observed. Hybrids have been found to be virtually sterile (Weimark *et al.* 1979).

1.3 *Impatiens glandulifera*

Impatiens glandulifera is an annual herb native to the Himalayas that is now widespread in Europe and it occurs in almost all regions of Britain (Beerling & Perrins 1993). It was introduced into Europe in 1839, and was brought into Britain as a garden ornamental, with the first records of it becoming naturalised in 1855 in Middlesex. The first record in the Tees, Tyne and Wear catchment area was in Durham city in only 1892 and it could not have spread so far in such a short time without the aid of humans (Perrins *et al.* 1993). It spread rapidly in the 19th century, being greatly aided by its popularity in gardens, although the rate of spread was greatest between the 1940s and 1970s. It occurs up to altitudes of 2,000-2,500 m in its native region, but has not been recorded above 210 m in Britain, and the limits to its northern distribution are determined by the length of the growing season (Beerling 1993). It is most commonly found in riparian habitats, where it is able to dominate large stretches of river, probably through all seeds germinating at the same time, thereby suppressing any competitors. This strategy works best in areas that have regular seasonal disturbances such as flooding in the case of riverbanks, which would explain the species success in such areas. *Impatiens glandulifera* can self-pollinate, so can produce selfed and outbred seeds. All plastids in the pollen grain are lost and so inheritance of chloroplast DNA is entirely maternal. *Impatiens glandulifera* is pollinated mainly by *Bombus lucorum* (Dunn 1977). In an observational study, this species comprised 91% of the visitors to the plant. *Apis mellifera, Bombus hortorum, Bombus pratorum and Bombus lapidarius, Bombus agrorum* were also observed, as were a number of Lepidopteran species (Dunn 1977). Bees can travel several kilometres and some species of Lepidoptera can travel hundreds of kilometres, which provides the possibility of large-scale gene flow. However, this would be expected to be rare, and although species of Lepidoptera were observed visiting the flowers of *I.*

glandulifera, it was not known for certain that they were feeding and hence taking part in the pollination.

Seeds have been found to have a dry mass of 2-35 mg (Beerling & Perrins 1993) and are dispersed up to 10 m by the capsule bursting open, although long distance dispersal occurs by water. The seeds float for a short time, prior to sinking and are then transported along riverbeds before germinating under water in a process called bythisolydrochory (Pysek & Prach 1993). The maximum dispersal distance has been estimated as 20 km (Wadsworth *et al.* 2000). It is not known whether seedlings are dragged onto riverbanks by wave currents or whether they become buoyant and then float onto riverbanks (Trewick & Wade 1986). Seedlings have a high light requirement for growth and may depend upon chance gaps forming or arriving in recently disturbed areas. Populations are often ephemeral as winter flooding can destroy them and make certain areas no longer suitable.

1.4 Invasions of the Study Species in Riparian Habitats

River systems act as biological corridors and aid downstream dispersal for a number of alien plants. Seasonal flooding causes disturbances that enable such aliens to become established. A comparison of the two species in riparian habitats in the Czech Republic found that *I. glandulifera* spread faster (Pysek & Prach 1993). *Impatiens glandulifera* was the species more associated with riparian habitats and its success as an invader was attributed to it producing a large number of seeds that are effectively dispersed. *Heracleum mantegazzianum* was found to spread down rivers following its initial invasion, but later it failed to continue to expand in such areas. It can disperse by wind and successfully compete with native plants in many different habitats. Therefore, it may later expand into other areas (which *I. glandulifera* does not tend to do) and this could account for its slow rate of spread in riparian habitats in the Czech Republic. In Britain however, its distribution more closely follows river systems and this may be because there are a higher proportion of areas adjacent to rivers that are unsuitable for colonisation, such as intensively farmed monocultures (Pysek 1991).

The species have different life history strategies. Being an annual, *I. glandulifera* produces a large number of seeds that germinate early and it is able to adapt to changing conditions because of its short life cycle. Both species can grow to

be over two metres in height, which helps them to outcompete natives and it may be that competitive ability is more important than life history strategy.

The mode of dispersal can also have an effect on the persistence of populations. Floating propagules, (such as seeds of *H. mantegazzianum*) are usually washed up onto riverbanks relatively close to flood limits. Therefore, populations become established well beyond normal river levels. Propagules that sink (such as the seeds of *I. glandulifera*) are dragged along the riverbed and are likely to reach riverbanks much closer to the normal water level of the river. The subsequent populations will thus become established much closer to the edge of rivers. Populations of species with floating propagules are therefore less vulnerable to floods than species with propagules that sink (Collingham *et al.* 2000).

1.5 Plant Population Genetics

The pattern of variation between populations depends upon a number of factors, including drift, selection, gene flow and demographic history. Hamrick and Godt (1996) reviewed studies of genetic variation in plants in order to find correlations between life history strategies, ecological traits and the level and distribution of variation. They found that population size, the breeding system of the plant and the scale at which a species was investigated all affected the variation found.

Comparisons of intraspecific genetic variation in plants and animals have shown that plant populations have significantly higher levels of variation than those of vertebrates (Hamrick 1982). The plant species found to have the greatest levels of variation within populations were widespread, perennial, wind pollinated outbreeders. Hamrick (1982) also noted that, depending on its level, gene flow between populations can have differing effects on variation in plants populations. Low levels of gene flow can introduce and maintain new alleles into a population so differentiating it from others, but high levels of gene flow can result in there being little difference in variation among populations as they are all well mixed. However, the efficiency of pollination can vary from year to year, so making any generalisations about gene flow tentative.

1.5.1 Population Size and Level of Fragmentation Between Populations

In species that can self-pollinate, such as *I. glandulifera*, the relative rates of selfing and outcrossing can vary even within populations and may result from genetic or environmental reasons. Dense populations have been shown to have lower rates of pollinator mediated outcrossing than less dense populations, probably because in such cases, individuals are often pollinated by close relatives (Turner *et al.* 1982).

Species with fragmented populations would be expected to have a higher proportion of overall variation due to between population variation than species with a more continuous distribution, as gene flow between fragmented populations would be lower. This was found to be the case for *Agave victoriae-reginae*, a perennial native to Mexico, found only on limestone outcrops such that populations are fragmented (Martínez-Palacios *et al.* 1999). Variation at ten allozyme loci were compared in ten populations, each composed of thousands of individuals. There were high levels of both within and between population variation. The within population variation was considered to be due to the large sizes of the population and the between population variation due to low levels of gene flow between isolated populations. The large population sizes would result in lower levels of gene flow between populations, as pollinators are less likely to visit more than one population when populations are large. Gene flow was also found to be very low in populations of the orchid *Goodyera procera* in Hong Kong (Wong & Sun 1999) as a result of geographical isolation between populations sampled.

1.5.2 Spatial Scale

The spatial scale over which variation is considered is an important factor when comparing genetic variation between populations (Hamrick & Godt 1996). Many studies have revealed evidence of isolation by distance. An example of this includes a study of microsatellite variation between cliff-top populations of *Beta vulgaris* (Raybould *et al.* 1998).

A comparison of genetic variation in endemic and widespread species found endemic species to have less overall diversity (Hamrick & Godt 1996). However, when considering variation at the same spatial scale, the endemic species had higher levels of variation. A study of variation in populations of the perennial *Lychnis viscaria* in Finland separated by several kilometres found high levels of between

population variation (Lammi *et al.* 1999). This contrasted with an earlier study of variation in the species which found relatively little variation between putative populations. However, in this case the populations were all found within an area smaller than the size of the home range of pollinators, and within the potential distance through which seed dispersal can occur. Therefore, levels of gene flow between populations would be expected to be much higher and hence between population variation much lower than in the study of populations several kilometres apart.

1.5.3 Breeding System

Consideration of the effect of breeding system on genetic variation found that species which are principally selfers have less overall genetic variation and very little variation within populations due to inbreeding (Hamrick & Godt 1996). The observed low levels of overall variation in selfing species could occur because many of these species are annuals or short-lived perennials with populations that vary greatly in number and so are vulnerable to extinction. New mutations are less likely to get fixed into populations or to spread to other populations than in outcrossing species. A comparison of four species of *Tricyrtis* sect. *Flavae* (Liliaceae) found that the species with the lowest levels of both within and between populations was the one selfing species (Maki *et al.* 1999).

The species found in the review by Hamrick and Godt (1996) to have the highest levels of overall diversity, and a high proportion of the overall variation within populations rather than between populations, were temperate, wind-pollinated conifers, which are relatively long-lived and disperse seeds mainly by wind. However, genetic variation differs widely between different species of pine, and this underpins an important finding of the review, which was that many of the observed patterns of genetic variation could not be explained solely by the life-history traits examined. This is because the evolutionary history and past distribution of a species also determine its genetic variation. For example, species that have recently been through a bottleneck or whose distribution had recently expanded from a small number of refugia may have less overall genetic variation than expected.

Another factor that may make predictions about genetic variation difficult is the influence of humans, which is expected to be important in invasive species. An example is a study of *Cordia alliodora*, which is a widespread timber tree. Chase *et al.*

(1995) studied genetic variation in population in the Atlantic and Pacific coasts of Central America. Genetic variation between populations was lower than for any other species studied over a similar range, and this is probably due to human introduction of the species into new areas. The low levels of variation in a species introduced for timber may also be due to selective breeding or the planting of selected varieties. These causes of lowered levels of variation would not affect accidental introductions.

The limited differences that were observed between populations of *C. alliodora* found in areas with different levels of rainfall were attributed to local micro-habitat adaptation resulting in different ecotypes. Adaptation to local conditions may occur by selection acting upon the seeds that land in a given micro-habitat. Evidence of selection acting on seeds was found in a study of the outcrossing tropical palm *Astrocaryum mexicanum* in which there was much greater between population genetic variation in adults than in seeds (Eguiarte *et al.* 1992). There were high levels of pollen movement between populations, which explained the similarity in seeds between populations, but only those adapted to the local conditions could survive into adulthood.

Low levels of within population genetic diversity is a common feature of many species of colonising plants, which is likely to be caused by small founding populations. Sun (1997) found this to be the case for three species of colonising orchids in Hong Kong with different breeding systems. There was also very little variation between populations of all three species, suggesting that the populations of each species in Hong Kong may come from a single founding population. The lack of genetic variation has not appeared to have had a detrimental effect on species persistence.

Keys and Smith (1994) studied variation in three widely spaced populations of *Prosopis velutina*, which had invaded grassland from neighbouring riparian areas. Subdivision was found in all populations and it was suggested that this arose because there was limited pollen and seed dispersal. This could have arisen initially when the species was confined to riparian areas. The linear nature of such habitats could have reduced the ability for gene flow and led to differentiation between populations.

A comparison of genetic variation between established and recently founded populations of invasive plants can be used to understand the mechanism of invasion. A study of recently established populations of the weedy perennial *Silene alba* found populations to have come from different source populations between which there had been very little mixing (McCauley *et al.* 1995). The genetic structure of new

populations depended upon the number of individuals that first colonise a new area and the populations from which they originated. Colonisation was found to most commonly occur by dispersal from the nearest existing population. The amount of variation between populations of invasive species depends upon the number of populations from which individuals colonising a new area are drawn. The study of genetic variation can therefore provide information about the origins, number and distribution of nascent foci.

1.6 Objectives

The aim of this project was to investigate genetic variation in invasive plants in order to determine the relative importance of anthropogenic introduction, dispersal and life-history strategies to their distribution. This was carried out by analysing the population genetic structure of *H. mantegazzianum* and *I. glandulifera* in the Tyne, Wear and Tees river catchments. High levels of dispersal and low numbers of introductions would be expected to lower genetic variation and vice versa.

The species are not thought to be able to disperse naturally between catchments and so independent introductions into each catchment must have occurred. Therefore, there may be large differences between catchments if these introductions came from different source material.

Within a catchment, if the distribution of each species was due to dispersal alone (and not multiple introductions), genetic variation between populations would be expected to be proportional to the distance along the rivers between them. As the species' distributions increased through dispersal, populations would become differentiated by genetic drift following founder events. Populations closer together would be expected to have larger levels of gene flow between them, which would have a homogenising effect. However, if there have been a number of independent introductions into each catchment, populations which were geographically close could be genetically very different, providing gene flow between the populations was low.

Populations of *I. glandulifera* are composed of larger numbers of individuals than those of *H. mantegazzianum*. *Impatiens glandulifera* is an annual with no viable seedbank. Therefore, there is likely to be more variation within populations of *I. glandulifera* than *H. mantegazzianum*.

Impatiens glandulifera has a greater dispersal ability than *H. mantegazzianum* and its pollinators have larger home ranges. Therefore, populations in the same catchment would be expected to be more similar than those of *H. mantegazzianum* and a higher proportion of the total variation in *I. glandulifera* would be due to between, rather than within catchment variation.

Considering the above, the hypotheses tested were:

1) Populations in the same catchment are more similar than populations in different catchments.
2) Within a catchment, there is a pattern of isolation by distance. Therefore, the pattern of variation is a reflection of the dispersal of the species.
3) *Impatiens glandulifera* has relatively more within population variation.
4) *Impatiens glandulifera* has a greater proportion of overall genetic variation due to between catchment (rather than within catchment) variation than *H. mantegazzianum.*

These hypotheses were examined using nuclear and chloroplast DNA microsatellite markers.

A model of the spread of the two species has been developed, called MIGRATE, which uses information about habitat types along rivers to predict the distribution of the species at regional and national scales (Collingham *et al.* 2000). In the Wear, predictions of the distribution of *I. glandulifera* were found to be broadly in accordance with the observed distribution. However, the model was less accurate for predicting populations of *H. mantegazzianum*. A suggested explanation for this was that the species has been present in the area for a shorter time and has yet to reach its potential distribution. It may be limited by its dispersal ability. The results of this study may help to improve the model by providing an estimate of gene flow and identifying possible sights of introduction.

Chapter Two

Introduction to the Measurement and Analysis of Genetic Variation

2.1 Molecular Markers and Measurement of Genetic Variation

There are a number of different molecular markers with which genetic variation can be measured, including RAPDs, AFLP, RFLP and microsatellites. In this study, a method was required that was fast to process (as large numbers of samples were analysed), gave reliable results and was sensitive enough for there to be variability within the relatively small study area.

Genetic variation within and between different populations of the species was analysed by looking at variable regions of chloroplast DNA (cpDNA), and nuclear and chloroplast microsatellites. Chloroplast DNA is less variable than nuclear genomic DNA, although intraspecific variability in the cpDNA of some other species has been found (Soltis *et al.* 1992). Microsatellite markers were chosen as they are highly variable and fast to process large numbers of samples. The development of microsatellite markers requires the screening of libraries and this was carried out for *H. mantegazzianum* and *I. glandulifera.*

2.1.1 Chloroplast DNA (cpDNA) Markers

Chloroplast DNA can be maternally, paternally or biparentally inherited, but in angiosperms, maternal inheritance is the most common (Mogensen 1996). Mechanisms preventing paternal inheritance include the loss of chloroplasts from generative or sperm cells, exclusion of male cytoplasm during gametic fusion and degradation of cpDNA from generative or sperm cells. There are examples of biparental inheritance, despite several of the above mechanisms being in place, in which there is leakage of male plastids into the egg cell. In species for which biparental plastid inheritance is the norm, male organelles reach the egg cell by plasmogamy. The reasons behind the different modes of inheritance are not known, although suggestions of their respective advantages have been made. Uniparental

inheritance may prevent any incompatibility between organelle genomes from arising, and maternal inheritance could have come about through mechanisms that prevent any foreign (and potentially dangerous) DNA from entering the egg.

The chloroplast genome is in complete linkage disequilibrium and is inherited as a single gene. Intraspecific cpDNA variation has been found in almost all detailed studies of a species' cpDNA and in some cases variation within populations has been found (Soltis *et al.* 1992). The cpDNA mutations found within a species are insertions, deletions and substitutions (which are usually detected when they occur in restriction sites).

Chloroplast inheritance in *I. glandulifera* is exclusively maternal (Beerling & Perrins 1993), and this is very likely to be the case for the *H. mantegazzianum*, although this has not been studied. Maternally inherited cpDNA can only disperse in seeds and therefore has less potential to disperse than nuclear DNA, which can be dispersed in pollen. Study of cpDNA variation can in such cases provide information about the relative importance of seed and pollen movement to overall dispersal and gene flow (McCauley 1995). For example, in cases in which pollen has a much greater ability to disperse than seeds, the study of spatially distinct populations would be expected to find greater variation between populations in cpDNA than in nuclear DNA.

2.1.2 Microsatellite Markers

Microsatellite DNA consists of simple sequence repeats, of a core one to five base pair unit, that occur in tandem. Variation in the number of repeats, usually a result of the gain or loss of a single repeat, is probably caused by DNA slippage (Levinson & Gutman 1987) although unequal crossing over may lead to larger changes in the number of repeats (Di Rienzo *et al.* 1994). The mutation rate is high, having been estimated at between 10^{-3} and 10^{-5} per generation (Queller *et al.* 1993). Microsatellites are evenly distributed throughout the genome, codominantly inherited, and highly polymorphic. These attributes make them very useful markers for studying intraspecific variation. They have significantly greater levels of polymorphism than allozymes, which have been classically used for analysing intraspecific variation (Lehmann *et al.* 1996). Microsatellites are abundant in mammals, which have hundreds of thousands of microsatellite loci, but have been found to be approximately five times less abundant in plants. A database search based largely on crop plants

found that microsatellites longer than 20 base pairs occurred on average once every 29 kb (Lagercrantz *et al.* 1993).

In order to find the most common microsatellites in plants and to select these as library probes, the Genbank nucleotide database was searched and the frequencies of different microsatellites compared. The two most common microsatellites were the dinucleotide repeats AT/TA and GA/CT and therefore these were used as probes to find microsatellite sequences in *H. mantegazzianum* and *I. glandulifera* as described in Chapter three.

The database search revealed, in agreement with other studies, that the relative abundance of various microsatellites differs between plants and vertebrates (Condit & Hubbell 1991; Akkaya *et al.* 1992; Chase *et al.* 1996). In vertebrates, CA/GT repeats are the most common sequences but they were found to comprise just 6 % of the microsatellites found in plants (Lagercrantz *et al.* 1993). An explanation for the high abundance of CA/GT repeats in animals has been that they may be specially maintained as being composed of purine-pyrimidine repeats, they could form Z-DNA (twisted in the opposite direction to ordinary DNA). AT/TA repeats, which were found to be common in plants, are also composed of purine-pyrimidine repeats. It has been proposed that the difference could have resulted by chance soon after the divergence of plants and animals (Moore, *et al.* 1991).

The relationship between the number of repeats of a microsatellite sequence and the number of alleles is not clear (Schlötterer 1997). However, many studies of plant microsatellites have found that longer sequences have greater numbers of alleles (Lagercrantz *et al.* 1993; Saghai Maroof *et al.* 1994; Chase *et al.* 1996) which may be due to a higher mutation rate. Four microsatellite loci found in natural populations of a tropical tree all had between five and 15 alleles, the locus with 15 alleles having the greatest number of repeats (Chase *et al.* 1996). In some cases, plant microsatellites have very high levels of polymorphism. Saghai Maroof *et al.* (1994) found very high variability in microsatellites in barley, with 37 alleles found at one locus. Microsatellites occur mainly in introns, but they also occur in exons, where their length appears to be limited, suggesting that they are under selection (Jarne & Lagoda 1996).

Repeats of different sequence motifs are often found together in clusters (Condit & Hubbell 1991; Chase *et al.* 1996) and although individual sequences may have ten or fewer repeats, the cluster as a whole may consist of up to a hundred repeats that can be considered together to produce a large number of alleles. Such clusters have been

termed compound microsatellites, with a single unbroken sequence being a pure microsatellite and a sequence broken by anomalous base pairs termed an interrupted microsatellite (Jarne & Lagoda 1996). The lengths of pure microsatellites are more variable than interrupted ones as the interruptions appear to stabilise the sequences.

Microsatellites usually have highly conserved flanking regions (Jarne & Lagoda 1996; Chee *et al.* 1996). When the sequences of these flanking regions are known for a species, they can be used as PCR primers for microsatellite analysis. In order to obtain these flanking sequences, a genomic DNA library of a species can be probed with labelled microsatellites and positive clones sequenced. Because microsatellites in plants are relatively infrequent, screening is much more efficient and a greater number of microsatellite sequences are likely to be found by screening a library enriched for microsatellites (Chase *et al.* 1996).

Variation at microsatellite loci among individuals can be measured by amplifying the loci by PCR using radioactively or fluorescently labelled dUTPs or primers and running the products on a polyacrylamide gel to separate out the products of different sizes.

Microsatellites also occur in the chloroplast and Weising and Gardener (1999) described ten universal chloroplast microsatellites, which were all mononucleotide repeats. These markers may be more variable than other regions of the chloroplast and could prove useful for making comparisons between seed dispersal and overall gene flow. Intrapopulation variation has been found using these loci (Drummond *et al.* 2000).

2.1.3 Assumptions for Analysis of Variation at Microsatellite Loci

There are three main assumptions that should be considered when analysing microsatellite variation. The first is that each locus is selectively neutral, and the second that the presence of null alleles is identified (Pemberton *et al.* 1995). The third is that for tests which pool variation at different loci, there must be verification of the independent assortment of the loci to ensure against linkage between markers, ie. the loci should be tested for linkage disequilibrium.

The observed variation at microsatellite loci occurs as a result of the effects of migration, mutation and genetic drift. Migration leads to a homogenisation of allele frequencies between spatially distinct populations whereas mutation and genetic drift can bring about the divergence of allele frequencies (Rousset 1997). The effect of

mutation on genetic variation may depend on its rate and mechanism as a high mutation rate with mutations both adding and subtracting repeat sequences could potentially reduce the overall gentic variation. Strong selective pressures can counteract these processes. If allele frequencies are stabilised by selective pressures, genetic distances between populations can be underestimated. Conversely, different selective pressures on linked loci in diverse areas can lead to different alleles becoming fixed in distinct populations and this will lead to an overestimation of genetic distance between populations. Microsatellite loci are predominantly neutral, but should not be assumed to be so as there are several examples of loci which can have deleterious effects on humans, most of which are trinucleotide repeats (Fu *et al.* 1991).

The presence of selection may be detected by the departure of genotype frequencies from the predictions of the Hardy-Weinberg equilibrium. The large numbers of alleles present at microsatellite loci and the often small- sample sizes can complicate the comparison of observed variation with the Hardy-Weinberg equilibrium. The tests which can be carried out include the comparison of observed and expected levels of heterozygosity (Edwards *et al.* 1992) and exact tests using conventional Monte Carlo or Markov Chain methods (Guo & Thompson 1992).

There are a number of possible causes of deviations from the Hardy-Weinberg equilibrium. Tests to determine whether there is a deviation from the Hardy-Weinberg equilibrium assume that there is random sampling of panmicitic individuals from a large population. This is often not the case for studies of natural populations (Robertson & Hill 1984).

An observed excess of heterozygotes may indicate the occurrence of disassortative mating or the presence of overdominant selection although this only occurs for the first generation following isolation and it only takes two generations for random mating to lead to the Hardy-Weinberg equilibrium to be re-established (Luikart *et al.* 1998). An excess of homozygotes may be due to a number of factors. There may be null alleles, caused by mutations in the primer binding sites such that an allele cannot be amplified by PCR (Callen *et al.* 1993). Selfing may be common in the population, which is the case with self-compatible plant species or the locus could be under directional selection. Another possibility is that populations may have substructure with random mating occurring only within the subpopulations. This is known as the Wahlund effect (Robertson & Hill 1984).

In order to determine which of the above explanations is relevant, additional information is required, such as knowledge of the breeding system of the species or population distribution. Examination of the inheritance of alleles, where possible, can allow for the identification of null alleles (Callen *et al.* 1993). *H. mantegazzianum* and *I. glandulifera* are self-compatible but the level of selfing is not known. Inbreeding can lead to an excess of homozygotes and levels of inbreeding can be estimated from the homozygote excess (Robertson & Hill 1984; Viard *et al.* 1997). The analysis of more loci will increase the ability to detect population substructure because each independent locus contains an independent history of the population depending on the amounts of random drift, mutation, and migration that have occurred.

2.2 Modelling Microsatellite Evolution

The first studies of genetic variation were carried out by assessing allozyme variation. New mutations in allozymes are thought to almost always give rise to completely new distinguishable alleles and the Infinite Allele Model (IAM) is based on this premise (Kimura & Crow 1964). The common methods used to determine the relationship between observed variation and population structure or genetic distance were developed based upon this model of the acquisition of variation. Since the process of mutation of microsatellites differs from the IAM, these methods may not be strictly applicable.

Microsatellites mutate by the process of slippage resulting in the gain or loss of one or more repeats, so new mutations are related in size to the alleles from which they originated. This also results in the total amount of variation being underestimated since alleles of the same size are not necessarily identical in terms of their evolutionary history. The Stepwise Mutation Model (SMM), which was first applied to the distribution of alleles observed from protein electrophoresis, has been used to better describe the process of microsatellite mutation (Pritchard & Feldman 1996, Weber & Wong 1993). Valdes *et al.* (1993) found that allele frequencies at 108 human microsatellite loci were consistent with the SMM, but were not able to conclude that the SMM exactly describes the mutation process because they used samples from more than one population.

In a study of the relationship between observations and computer simulations based on the SMM, Shriver *et al.* (1993) compared the number of alleles, the range of allele sizes and the number of modes in the distribution of alleles. This study avoided the effects of population substructure by examining loci from large homogenous populations, and used loci with less than 50% heterozygosity, so that expectations of the IAM and the SMM could be distinguished. The markers studied were three classes of variable numbers of tandem repeats (VNTR's) including 31 microsatellites with 1-2 bp repeats, 12 microsatellites with 3-5 bp repeats, and 11 minisatellites (15-70 bp repeats). The comparisons indicated that all the microsatellites with 3-5 bp repeats but only 65% of microsatellites with 1-2 bp repeats matched the simulation values at all three measures. The 1-2 bp microsatellites had a greater number of alleles and a greater allelic size range than predicted by the SMM. A possible explanation for these deviations is that in addition to slippage, a low frequency of multi-step mutations also occurs. Di Rienzo *et al.* (1994) developed a model, based on coalescence theory, that more accurately explains the observed variation at dinucleotide repeat microsatellites. This model, the Two Phase Model (TPM), integrates the mutational process of the SMM with the possibility for mutations of a larger magnitude. The two models were applied to 10 dinucleotide repeat microsatellites genotyped in a well defined human population. Levels of heterozygosity and the frequency of the most common allele produced by the simulation were compared to the observed values. In eight of the ten loci the SMM did not apply but the TPM predictions were accurate. The multi-step mutations are thought to be caused by slippage through stem-loop formation, or by unequal crossing over (Di Rienzo *et al.* 1994). Although further work is needed to fully understand the mutational processes, the above results suggest that different classes of microsatellites may vary in their mode of evolution. VNTRs that evolve following the IAM will have the lowest levels of homoplasmy (alleles of the same size resulting from reasons other than inheritance from a shared ancestry) and therefore their use will overcome problems of the underestimation of variation encountered with loci that follow the SMM. Homoplasmy can occur via the SMM when a set of repeats are gained and then lost, resulting in no overall change in allele size.

2.3 Analysis of Genetic Variation

There are a number of statistical tests that can be used to analyse genetic variation at the population level. The amount of differentiation between populations can be quantified in terms of F_{ST}. This parameter can be estimated from population comparisons of the proportion of variance that accounts for between as opposed to within population differences (Nei 1973; Weir & Cockerham 1984) and can be defined as:

$$F_{ST} = (S_t - S_w)/S_t \qquad (1)$$

where S_w and S_t are estimated as the within population and total variances respectively. F_{ST} does not consider different mutational relationships among alleles.

Slatkin (1995) proposed the use of a new statistic, R_{ST}, which is analogous to F_{ST} but was designed to fit a generalised stepwise mutation model which is similar to the TPM. R_{ST} is defined as

$$R_{ST} = (S_t - S_w)/S_t \qquad (2)$$

where S_w is twice the average of the estimated variances in allele size within each population. S_t is twice the estimated variance in allele size in all populations. The estimates of variance were obtained using unbiased estimators. Simulations showed that R_{ST} usually provides a less biased estimator of population structure than F_{ST}. More than one locus can be examined at the same time by finding the weighted average values of among population variance and total variance across loci (Goldstein 1995a).

Another measure of genetic distance is $(\delta\mu)^2$ which, like R_{ST} is a mutation-based distance and often gives similar results (Perez Lezaun *et al.* 1997). $(\delta\mu)^2$ is the sum of the square of the difference between the average repeat size in two populations. This can be seen with the equation

$$(\delta\mu)^2 = (m_x - m_y)^2 \qquad (3)$$

where m_x and m_y are mean repeat numbers at each locus in populations x and y. If there is reproductive isolation and mutational-drift equilibrium, and stepwise mutation, then $(\delta\mu)^2$ is a linear function of the separation time between populations (Goldstein *et al.* 1995b).

Considering the relatively short time that *H. mantegazzianum* and *I. glandulifera* have been present in Northeast England and that populations may be ephemeral, there is unlikely to be an equilibrium between migration and genetic drift. This equilibrium

is an assumption of the island model, on which some of the relevant tests are based. Assignment tests do not require such assumptions and are based on multilocus data and so may be particularly useful when considering microsatellite variation in such populations (Paetkau *et al.* 1995; Davies *et al.* 1999). Assignment tests calculate the likelihood of drawing a particular multilocus genotype from a number of potential populations. The more loci included, the more accurate the test is likely to be (Boeklen & Howard 1997).

2.3.1 Estimation of Gene Flow

The quantification of dispersal in plants often proves difficult, especially considering long-distance dispersal. Direct measurements may prove inaccurate because long distance dispersal is usually very rare but is of great significance to genetic variation, colonisation of new areas and metapopulation structure (Silvertown 1991). Indirect measurements of gene flow can be obtained from genetic markers by assessing the distributions of alleles. The factors which affect this distribution include seed and pollen migration and random genetic drift (Govindaraju 1988). If dispersal is considered to refer to the dispersal of seeds and gene flow the migration of both seeds and pollen, different types of molecular markers can be used for indirect estimates of dispersal and gene flow. Gene flow will only occur through seed dispersal in chloroplast markers, which are maternally inherited in most angiosperms, and not in pollen. Therefore the comparison of estimates of gene flow from nuclear and chloroplast markers can give a measure of the relative importance of pollen versus seed migration.

In order to produce indirect estimates of gene flow from parameters of genetic variation, a demographic model is required to describe the way in which dispersal occurs, of which there are a number. The most commonly used model for derivations of F_{ST} is the infinite island model. This assumes that there are an infinite number of populations consisting of equal numbers of individuals which each give and receive an equal number of migrants to every population.

There are a number of other demographic models, many of which are more realistic than the infinite island model (Cockerham & Weir 1993). These include:-
one and two dimensional stepping stone models, which assume a constant migration rate with migration occurring from one step to the next adjacent one (Tufto *et al.* 1996); continuum models in which the migration rate is a function of distance (Slatkin

& Barton 1989) and migration matrix models, in which migration rates can vary and are determined for each pair of populations in a matrix (Bodmer & Cavalli-Sforza 1967).

Studies of allozyme variation have derived estimates of gene flow from F_{ST} using the equation:

$$F_{ST} = 1/(1 + 4N_e m) \quad (4)$$

Where N_e is the effective population size and m the proportion of immigrants in a population so that $N_e m$ is the number of migrants (Wright 1978). This equation is based on the assumptions of Wright's isand model, that levels of gene flow and genetic drift are at equilibrium, populations are of constant size, all populations contrbute equally towards migrants and migration occurs at random. If this is the case then the rate of gene flow will be inversely related to the differentiation between populations. However, in natural situations, many if not all assumptions are likely to be violated. In the case of riparian species with linear water-mediated dispersal patterns, it is very likely that downstream dispersal will not occur equally from all populations. Violation of the assumption of uniform migration has been found to lead to an underestimate of genetic variation and thus an overestimate of gene flow (Giles & Goudet 1997). Equilibrium may take tens of generations to be reached. *Heracleum mantegazzianum* is a perennial which can take four years to flower, which first occurred in the Tyne in the 1920s but was not recorded in the Tees or Wear until the 1940s and so it is very likely that many local populations would have been present for less than ten generations. Populations of *I. glandulifera* may be subject to metapopulation dynamics (see below) due to their vulnerability to extinction and the large potential for downstream dispersal. Where populations resemble a metapopulation, there can be large departures in variance from that predicted from the island model, on which the model of dispersal is based (Whitlock & McCauley 1990; Giles & Goudet 1997). Violations of the assumptions may lead to imprecise estimates of gene flow from F_{ST} values but are likely to be correct within a few orders of magnitude (Whitlock & McCauley 1999).

2.3.2 Metapopulation Dynamics

Many of the tests used to analyse genetic variation are based on the assumption that populations have reached a balance between migration and genetic drift. However, when considering the spread of a species into a new area, especially into habitats

vulnerable to environmental extremes, it is unlikely that such a balance will have been reached.

Metapopulations are composed of demes that are subject to frequent extinctions followed by frequent recolonisations. The metapopulation as a whole is therefore able to persist for longer than the individual demes of which it is comprised. The pattern of genetic variation found within a metapopulation is determined by a number of factors. These include the way in which groups of colonists are formed and the proportion of colonists entering new areas, relative to those going extinct (Wade & McCauley 1988). Although newly formed demes would be expected to be related to the demes from which they arose, genetic drift following founder events can complicate this. The number of individuals within demes and migration rates can vary because of floods and other stochastic events (Whitlock 1992). In such cases, models which assume constant migration rates could prove misleading.

McCauley *et al.* (1995) looked at F_{ST} amongst populations of *Silene alba* of varying age. *S. alba* is a dioecious ruderal plant which is often subject to metapopulation dynamics. F_{ST} values were greater among the more recently founded populations. Where rates of extinction and recolonisation and therefore gene flow were higher, F_{ST} values were higher. Therefore, estimates of gene flow would be lower in systems in which the actual rates of gene flow are higher. Similar disparities were found by Giles & Goudet (1997), so exposing the unsuitability of metapopulations for indirect estimates of gene flow. However, Whitlock and McCauley (1990) showed that the common origin of colonisers could be determined by an indirect approach. A model was developed in which the F_{ST} value between newly colonised sites (F_{STo}) could be expressed as a function of the F_{ST} value between older populations, the number of colonisers, and a parameter, Φ, which is determined by the pattern of colonisation. The number of colonisers could be estimated because McCauley *et al.* (1995) were able to use census data and therefore obtain an estimate of Φ. Whitlock and McCauley (1990) defined two extreme patterns of migration, the migrant pool mode and the propagule pool mode. In the migrant pool mode, colonisers are drawn at random from all possible populations whereas in the propagule pool mode, all colonisers are drawn from the same source population. Φ will be 0 in the case of the migrant pool mode and 1 in the case of the propagule pool mode. The size of Φ can thus be used to estimate the mode of colonisation. The estimates of Φ obtained by McCauley *et al.* (1995) were 0.73 for allozymes and 0.89 for cpDNA markers. This suggests that most founders are drawn from the same source population.

In order to discover the effects of population size and distance between population on gene flow in *S. alba*, Richards *et al.* (1999) set up experimental arrays to enable the direct measurement of gene flow. Two sets of eight experimental populations were set up in a line at intervals of 20 m and 80 m. The populations were set up to be alternately homozygous for alleles at specific allozyme loci so that gene flow between the different populations could be identified. The populations consisted of different numbers of individuals in order for the effect of population size to be studied. Populations separated by 20 m had significantly greater levels of gene flow than those separated by 80 m, as may have been expected. However, there was considerable variation in the levels of gene flow in populations separated by the same distances, which may have been caused by differences in the number of flowers at the nearest populations when gene flow occurs. Population size can affect pollinator foraging behaviour as pollinators are likely to visit more flowers within a large population and therefore pollinators would be less likely to visit other populations. This was found to be the case for the experimental arrays separated by 20 m where the smallest populations had the highest levels of gene flow. However, in the arrays where populations were separated by 80 m, the smallest populations showed no gene flow at all between populations. Therefore, it is the interaction between population size and distance between populations that appears to determine the level of gene flow between populations following colonisation, and so immigration rates may be highly variable. A small population arising far from the nearest source population may not be visited by pollinators from other populations at first so levels of gene flow would be very low. However, if the population continued to grow, eventually, it would be more likely to be visited by pollinators and so levels of gene flow would rise. If the population became very large, pollinators would not need to then visit other populations, so levels of gene flow would then decrease. Gene flow can therefore vary temporally following colonisation.

Demes within a metapopulation may be subject to large fluctuations in numbers of individuals. When numbers drop, and there is little gene flow into the area, this may result in a bottleneck. Luikart *et al.* (1998) showed how distortion in allele frequency distributions could be used to identify recent bottlenecks. The distortion occurs because the number of rare alleles is likely to be reduced, because some will be eliminated from the population, and this results in a shift in the distribution of allele frequencies. This distortion was expected to last for not more than a dozen generations. Considering that *H. mantegazzianum* has only been present in the study

area for 72 years and that individuals take up to five years to set seed, most bottlenecks since entering the study area could be detectable.

2.3.3 Effect of Inbreeding on Genetic Variation

Metapopulation dynamics have been shown to affect genetic variation between populations (section 2.3.2). Another factor that may affect genetic variation in populations of the study species is their self-compatibility. Inbreeding populations have lower levels of heterozygosity than outbreeding populations, although the level of variation between populations may depend upon the level of inbreeding. Variation within populations would be expected to be lower than in outcrossing species since the effective population size is reduced by inbreeding.

Viard *et al.* (1997) investigated the effect of inbreeding on genetic variation in subdivided populations by looking at populations of the hermaphrodite freshwater snail *Bulinus truncatus* in West Africa and the Mediterranean. Populations of *B. truncatus* are discrete and patchily distributed and annual flooding and drought leads to frequent variation in population sizes and a number of extinctions and recolonisations. Therefore, this species provides a good example for which to look at the effect of metapopulation dynamics and self-fertilisation on genetic variation between populations. Genetic variation in 38 populations, at distances of between less than a kilometre to thousands of kilometres apart, was measured using four microsatellite loci.

The study found low levels of heterozygosity in almost all populations and levels of selfing were estimated at over 80%. Levels of variation within populations were highly variable, with four populations being monomorphic at all four loci, whereas one population of 37 individuals had 19 alleles at one locus. Considerable genetic variation occurred among populations at all spatial scales, which was thought to be a consequence of the high levels of selfing. The effects of genetic drift and gene flow on genetic variation depended on the spatial scale examined. Water levels can vary greatly and if a pool dries out, in the following rainy season, it may be recolonised from a number of surrounding pools. The high levels of within population variation found in some populations was thought to be due to recolonisation from a number of surrounding populations following a local extinction.

A study of populations of the self-compatible orchid *Goodyera procera* in Hong Kong found no significant correlation between population size and genetic variation

within populations (Wong & Sun 1999). It was suggested that this could have been due to inbreeding lowering variation within populations regardless of their size and the species may also have been through a bottleneck.

Low variation within a population could also result from extinction followed by recolonisation by a small number of individuals, genetic drift following recent recolonisation and cyclical variation in population sizes which cause populations to go through bottlenecks. The effective population size is lower in selfing species and migration rates are lower than in outbreeding species and the effects of genetic drift are enhanced.

2.3.4 Seed Versus Pollen Flow

Gene flow between plant populations can occur by dispersal of pollen and dispersal of seed. In most species of plants, pollen dispersal, which can occur by wind, insects, birds or mammals (eg. bats), is greater than the potential for seed dispersal, which occurs by mammals, wind, mechanical ejection or water. However, in the case of *H. mantegazzianum* and *I. glandulifera*, the potential for seed dispersal is far greater than that for pollen dispersal, with maximum dispersal distances for a single propagule estimated as 10 km and 20 km respectively. In contrast, the estimated maximum pollen dispersal distances are 1 km and 5 km respectively (Willis 1999).

In order to completely describe gene flow in plants, both methods of gene flow must be taken into account. Methods for the indirect estimation of gene flow have been described above. The vast majority of direct estimates of gene flow have been carried out only for gene flow via pollen and not by seed. When paternity exclusion analysis is possible, foreign pollen can be identified (Ennos 1994). The migration rate of pollen can be found by consideration of the proportion of foreign pollen in the total amount of pollen.

In recent years, molecular markers in organelle genomes have been used in population studies. These markers differ from nuclear markers because they are commonly inherited in a uniparental fashion. Chloroplast genomes are predominantly inherited maternally in angiosperms (section 2.1.1). The difference in the mode of inheritance is expected to lead to differences in gene flow between the different classes of markers. It would be expected that levels of gene flow in maternally inherited markers would be lower than that for biparentally inherited markers since these can disperse in pollen as well as in seeds (Birky *et al.* 1994). The extent of these

differences will be determined by the relative amounts of seed and pollen flow. In the case of *H. mantegazzianum* and *I. glandulifera* this difference would be expected to be lower than for other angiosperms because they both produce seeds capable of long distance dispersal.

Ennos (1994) produced a model to relate levels of gene flow for nuclear, maternally inherited and paternally inherited markers to levels of pollen and seed flow between populations. The model had a number of assumptions, including that the plant species is diploid and hermaphrodite (as is the case with the two afformentioned study species), and is distributed according to the infinite island model with a constant population size and level of gene flow. In the case of *H. mantegazzianum* and *I. glandulifera*, these conditions are unlikely to have been met. Assuming an equilibrium between drift and migration, it was shown that for maternally inherited markers with a migration rate m_m, $F_{ST\,(m)}$ is given by:

$$F_{ST\,(m)} = 1/(2N_m + 1) \qquad (5)$$

This can be contrasted to the equation (4) for nuclear markers. When the model was modified to allow for inbreeding, which will occur in all self- fertilising plants, the effective population size is reduced in relation to the extent of inbreeding but the number of alleles in a given population does not vary. It is necessary to consider the level of inbreeding in order to avoid levels of *Nm* being underestimated. In addition, it was shown that the relative rates of pollen and seed migration between populations can be estimated from a comparison of F_{ST} values for nuclear and maternally inherited markers. However, this relationship is only valid when there are low levels of absolute pollen or seed migration. This may not be the case for the study species.

2.3.5 Phylogenetic Tree Construction

Cavalli-Sforza and Edwards (1967) established a method for constructing trees from data on genetic variation between populations (which can also be used at the level of individuals or species). All populations are considered to be independent, and are taken to diverge in the pattern of a branching random walk in a Brownian motion process. Splitting occurs at random and different tree forms are tested until that which best fits the data is found (Cavalli-Sforza & Edwards 1967). There are a number of different modes of splitting and the random walk can have different properties resulting in different methods that may be used.

Fitch and Margoliash (1967) developed distance methods for tree construction, with genetic distance matrices of pairwise population comparisons providing the raw data. The "additive tree" method assumes that distances are equal to the sum of branch lengths between populations and branch lengths are unconstrained. The "ultrametric tree" method is based on the "additive tree" model but also assumes that there is a molecular clock. Therefore, branches of the tree are constrained such that the total length from the root of the tree to any population is the same. Saitou and Nei (1987) developed a neighbour-joining method which constructs trees by the successive clustering of populations, with branch lengths fixed as the populations are added. The method starts with all populations at the end of one branch originating from the same point, and then groups pairs of branches in succession, so building up a tree. Branch lengths are determined by the genetic distances between pairs of populations, and the best tree is that with the lowest overall total of all branch lengths.

A maximum likelihood approach can also be used to construct trees, where the alleles present at each locus in all populations are used instead of a genetic distance matrix, which is used in the methods described above. Felsenstein (1981) introduced a restricted maximum likelihood approach whereby the differences between populations are considered, which avoids problems caused by nuisance parameters that arise from each new character considered in the analysis. Trees can be constructed to find the maximum of the restricted likelihood using and an algorithm for searching between different tree topologies and within a topology using a combination of pruning and the pulley principle. Pruning involves removing one pair of populations at a time and considers the likelihood without them. The pulley principle alters branch lengths since only relative differences are considered and therefore the lengths of two branches with a common node can be varied in tandem.

The trees produced by the above methods are unrooted and the branch lengths are best thought of as expected distances, rather than actual path lengths (Felsenstein 1984).

Chapter Three

Methods

3.1 Collection of Leaf Material

Samples of the three species were collected from the Tyne, Wear and Tees river catchments. The Environment Agency carried out a five year River Corridor Survey of the entire area of the above three catchments and divided all rivers and tributaries into 500m long sections. The species present were noted and maps of the sections were drawn. These maps were used to identify the locations of the three species. Section 1 of each catchment was given to be located at the mouth of the river, so for example, population Tees 152 is found 76 km upstream from the mouth of the Tees. Tributaries were numbered similarly.

Since it was possible that there would be little variation in the study area, initially samples were collected from the most spatially distinct sections of the three catchments. Therefore in each catchment, where possible, each species was sampled at three points along the main river and at the most upstream 500 m stretch of four spatially distinct tributaries. *Impatiens glandulifera* is relatively abundant in all tributaries, but *H. mantegazzianum* is found along just one stretch of the Wear and in only one tributary of the Tees and Tyne. Populations of both species, but particularly of *I. glandulifera* are ephemeral and at a number of locations, the species was not found despite being recorded between 1990-1991 in the River Corridor Survey. Although plants were collected from a number of locations, initially, two populations were genotyped from each catchment. The observed pattern of variation found was then used to determine which further populations should be genotyped. One population of *H. mantegazzianum* and two of *I. glandulifera* were resampled two years later. The distribution of the species in the study area and the locations of populations genotyped are shown in Figures 3.1 and 3.2.

Where there were fewer than 30 individuals in one 500 m reach, leaves from all individuals present were collected. At sites along tributaries, or the most upstream of a major river, with more than 30 individuals, 30 leaves were collected at random from the most upstream clump of plants numbering more than 30. At the most downstream sites, leaves were collected from 30 individuals in the most downstream

clump and at midstream sites from the most midstream clump. Clumps were divided into a 10 X 10 grid. Thirty randomly generated co-ordinates, corresponding to thirty grid cells, were obtained and the first plant encountered in each grid cell was harvested. The approximate number of individuals of each species present in each 500 m reach was noted, as was the presence of the native *Heracleum sphondyllium* (section 1.2.1). The above information for each population at which leaves were sampled is given in Table 3.1.

The method of collection can greatly affect the quality and yield of DNA that can be extracted (Nickrent 1994). Leaf material was used as it is the easiest material to homogenise. One leaf was collected from each plant of *H. mantegazzianum* and three from each individual of *I. glandulifera.* Harvested leaves from each plant were placed in plastic bags and kept on ice until they could be stored at -80^{o}C (Mason-Gamer *et al.* 1995). The precise areas of collection at each site were mapped and the number of individuals was estimated.

Figure 3.1 Distribution of *H. mantegazzianum* in the study area

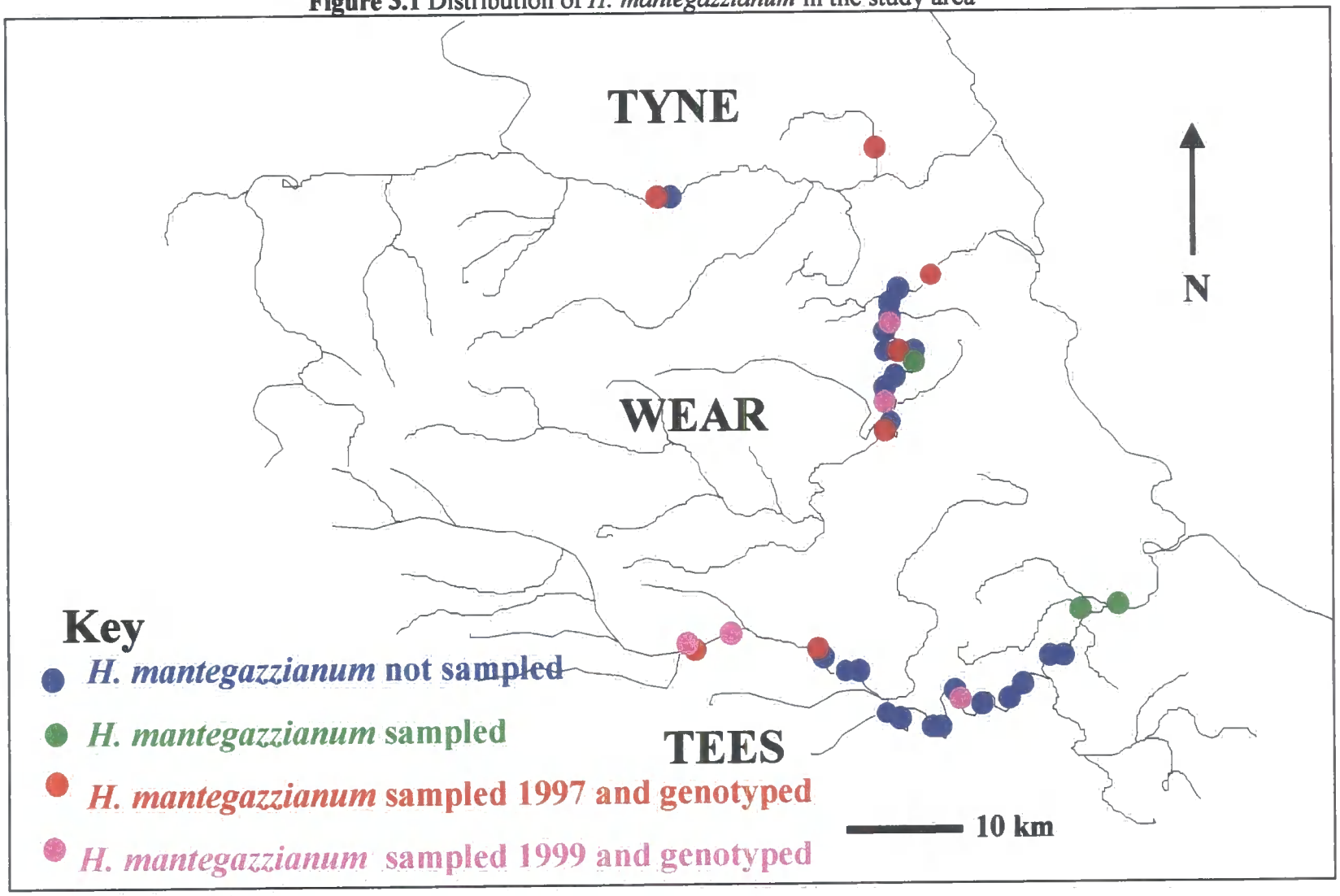


Figure 3.2 Distribution of *I. glandulifera* in the study area

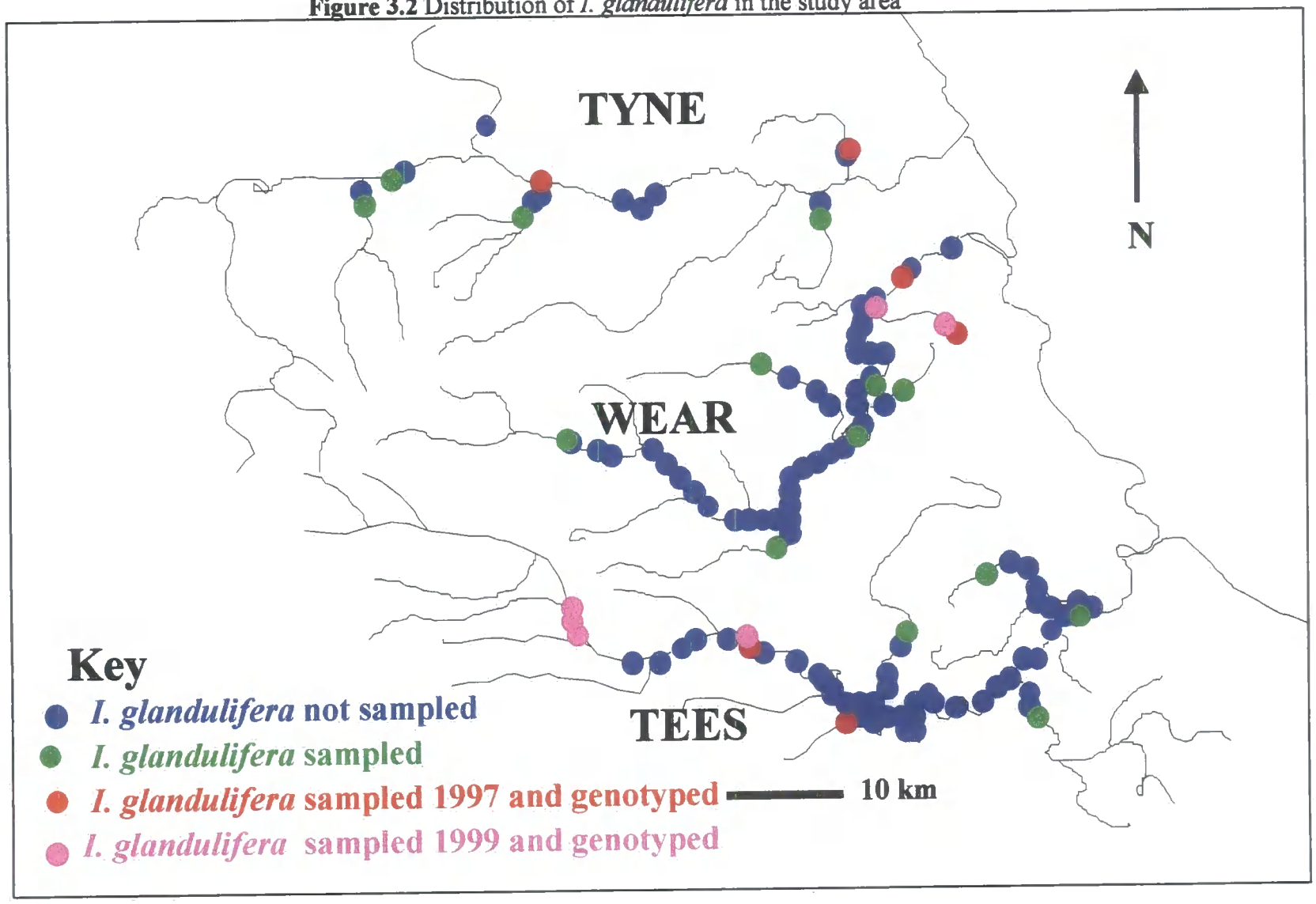


Table 3.1 Sites at which populations of *H. mantegazzianum* (H) and *I. glandulifera* (I) were sampled. The numbers given were the estimated total numbers of individuals present at each location.

Site	Approx. # H	Approx. # I	H. sphondyllium	Bank	Distance from river (m)	Genotyped
Tees, Lev 66		60		L	0	**X**
Tees, Lus Beck 3	40			L,R	0	**X**
Tees, Lus Beck 6		200		L,R	0	**X**
Tees, Skerne 58		40		R	1	**X**
Tees, Spa 3		40		L,R		√
Tees 20	100	200	√	L,R (I)	0	**X**
Tees 59	28		√	L	5	√
Tees 60		100		L,R	0	**X**
Tees 140		200		L		√
Tees 152	40		√	L		√
Tees 162	60			L,R		√
Tees 186		50		L	3	√
Tees 188		35		L	0	√
Tees 188		1000		L	0	√
Wear, Browney 52		50		L	<1	**X**
Wear, Rainton 1		200		L,R		√
Wear, Rainton 52		50		L,R	0	√
Wear 18	29	70		R	3	√
Wear 35	50		√	R		√
Wear 46	24		√	R	3	√
Wear 51	2	500	√	R,L(I)	2(H), 1 (I)	**X**
Wear 71	100		√	R		√
Wear 72	10	500		L,R(I)	3	**X**
Wear 77	200			L		√
Wear 89		200		L,R		**X**
Wear 179		500		L	10	**X**
Tyne, Devil's 2		150		R	1	**X**
Tyne, Don 4		50		L,R		**X**
Tyne, Ouse 4	40	100		L,R	<1	√(H, I)
Tyne, Tip Burn 2		300		L,R		**X**
Tyne 70		200	√	L	3	√
Tyne 92		200		R	2	√
Tyne 94	26	100	√	R	0	√(H)
Tyne 161		200		R		**X**

3.2 Molecular Biology Methods

All chemicals were obtained from Sigma or BDH and were of analytical grade or higher unless otherwise mentioned. Sequencing of DNA and the running of GeneScan™ gels for the sizing of microsatellite loci were carried out by the DNA sequencing laboratory at Durham University using an ABI 377A DNA sequencing machine.

3.2.1 CTAB isolation of Plant DNA- modification of Murray and Thompson (1980)

Extraction of DNA from plants can prove difficult because of the presence of secondary compounds such as polyphenolics and inherent enzymes that can degrade DNA, and also because of problems removing polysaccharides (Milligan 1997). Hexadecyltrimethylammonium bromide (CTAB) is a cationic detergent that is often used in plant DNA isolation as it solubilises plant membranes and forms a complex with DNA.

Ten to 20 g of leaf tissue was frozen in liquid nitrogen and ground with acid-washed sand in a pestle and mortar. The powder was transferred to a 50 ml centrifuge tube and an equal volume of 2 X CTAB extraction buffer (100mM Tris-HCl pH 8.0, 1.4 M NaCl, 20mM EDTA (ethylenediaminetetraacetic acid), 2 % CTAB), preheated to 60°C, was added. The mixture was incubated for 30 minutes at 60°C, allowed to cool, extracted with an equal volume of chloroform/isoamyl alcohol (24:1), and centrifuged at 1,500 g for 10 minutes. The aqueous phase was extracted and a $1/10^{th}$ volume of 10% CTAB buffer (10% CTAB, 0.7M NaCl) was added to the resultant removed aqueous phase. The aqueous phase was extracted as previously with chloroform/isoamyl alcohol. The DNA/CTAB complex was precipitated with two volumes of CTAB precipitation buffer (50mM Tris-HCl pH 8.0, 10mM EDTA pH 8.0, 1% CTAB). The mixture was incubated overnight at 20 °C. Originally, precipitation was carried out for one hour, but it was found that yields were vastly improved by precipitating overnight. The DNA was pelleted by centrifugation at 1,500 g for 5 minutes, resuspended in 450 µl of 1 M NaCl and transferred to a 1.5 ml Eppindorf tube. The DNA was precipitated by the addition of 900 µl of room temperature 100% ethanol. Following centrifugation at 12,000 g for 15 minutes in a Beckmann microcentrifuge, the pellet was resuspended in a 20 µg/ml solution of

RNase A and incubated at 37 °C for 10 minutes. The solution was diluted in one volume of TE and a phenol/chloroform extraction was carried out (section 3.2.2).

3.2.2 Phenol/Chloroform Extraction of DNA

DNA was taken up in 200 µl TE and an equal volume (200 µl) of phenol/chloroform/isoamyl alcohol (25:24:1) was added, mixed gently, left to stand for a few minutes and centrifuged for one minute at 12,000 g in a Beckmann microcentrifuge. The aqueous phase was collected and 200 µl chloroform/isoamyl alcohol (24:1) was added. The solution was mixed and the aqueous phase was collected. Ten µl of 3M sodium acetate was added, mixed and then 500 µl of cold 100% ethanol was added. The solution was mixed gently and placed in a freezer for one hour. The sample was then centrifuged for 15minutes and the supernatant discarded. The pellet was washed in 80% cold ethanol and mixed. The sample was left for 10 minutes and centrifuged for two minutes. The supernatant was discarded and the pellet dried under vacuum. The pellet was dissolved in 100 µl TE and the sample stored at –20 °C.

3.2.3 Extraction of DNA for PCR

The above extraction methods gave the highest yields of DNA, but for the extraction of DNA from the numerous leaf samples required for screening, a faster, if less efficient method was preferable. It is important when carrying out a large number of extractions that there is no cross contamination. The step most likely to introduce cross contamination is the grinding which is necessary to break down the cellulose cell walls and was used in all the above methods. In order to avoid problems of cross-contamination during the grinding step, leaf tissue was ground in a 1.5 ml tube using a blue-tip that had been melted to form a pestle.

3.2.4 Small-scale Extraction of DNA for PCR

Extraction of DNA of *H. mantegazzianum* and *I. glandulifera* for amplification of microsatellites was carried out using a small-scale version of the method given in section 3.2.1. All centrifugations were carried out in a Beckmann microcentrifuge at 12,000 g. Two hundred µl of 2 X CTAB extraction buffer (100mM Tris-HCl pH 8.0,

1.4 M NaCl, 20mM EDTA, 2 % CTAB) was pipetted into a 1.5 ml tube along with a pinch of sand and the tube heated to 65 °C. Ten to 50 mg of frozen leaf tissue was placed into the 1.5 ml tube and ground up using a 1ml pipette tip whose sharp end had been melted in a flame and moulded into a pestle by placing it inside a 1.5 ml tube. The total volume was brought to 500 µl by the addition of 1 X CTAB extraction buffer. The mixture was incubated at 65 °C for two hours, allowed to cool and then extracted with a 24:1 mixture of chloroform/isoamyl alcohol. Following centrifugation for two minutes, the supernatant was collected and 50 µl of a solution of 10% CTAB and 0.7 M NaCl was added. The aqueous phase was extracted as previously with chloroform/isoamyl alcohol. The DNA/CTAB complex was precipitated with 1000 µl of CTAB precipitation buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 1% CTAB). The mixture was incubated overnight at 20-25 °C and then centrifuged for five minutes. The pellet was rehydrated in 200 µl of resuspension buffer (10 mM Tris, pH 8.0; 1 mM EDTA, pH 8.0; 1 M NaCl). The DNA was precipitated by the addition of 400 µl of cold 100% ethanol. Following centrifugation for 15 minutes, the pellet was washed for five minutes, in 80 % ethanol, and centrifuged for five minutes. The pellet was resuspended in 50 µl TE.

3.3 Design of Chloroplast DNA Markers

The "Genbank" database has the entire chloroplast DNA sequence of a number of taxonomically distinct species. The chloroplast sequences of five species, which are as distinct from each other as from the study species, were searched and corresponding introns found. These sequences, including their surrounding conserved regions, were compared by aligning them using the "Genetics Computer Group" (GCG) software package. The most variable regions could be identified as well as the most highly conserved regions. Areas suitable for sequence analysis have 300-600 base pairs which are highly variable among the five species, with flanking regions of over 17 base pairs that are almost identical. Possible highly conserved regions for use as primers were also located from previous studies. Demesure *et al.* (1995) listed universal cpDNA primers that amplified regions of over 3,000 base pairs, for analysis by restriction fragment length polymorphism (RFLP). These regions are too long for sequencing, but regions up to 700 base pairs either side of the given primers were

compared amongst the five species. Taberlet *et al.* (1991) listed three sets of primers that amplify regions of 300 to 800 base pairs. Several of the published primers amplified regions that had already been identified by the database search. Twelve possible variable regions were found, and their suitability and compatibility as polymerase chain reaction (PCR) primers were tested using the "oligo" and "Amplify" software packages. This was necessary as primers must fulfil a number of criteria: each primer must be of the correct length (17-24 base pairs) and have ends that do not cause it to tend to fold over on itself in a hairpin or form dimers with itself. In order to be compatible, two primers must have similar annealing temperatures during the PCR reaction and must not be able to form dimers by the 3' ends attaching to each other.

All but one of the published sets of primers was found to have a tendency to form dimers. However, with slight modifications, such as cutting out a few base pairs on one side of a primer and adding some at the other side, two suitable sets of primers were found (Table 3.2). One set of primers amplified a region in between the published universal primers situated in the exons tRNA-Ser(UGA) and psII 44kd protein (Demesure *et al.* 1995), using a modification of the tRNA Ser(UGA) primer and a second primer designed in a conserved region that follows an intron of about 430 base pairs. The second set of primers amplify a region in-between tRNA L(UAA) and tRNA F(GAA) which was amplified by Taberlet *et al.* (1991), the differences being that the primer in exon tRNA L is two base pairs shorter at the 5' end than that published and the primer situated in tRNA F exon is found 19 base pairs closer to the 5' end than that published.

Intraspecific variation has been found in the same regions of cpDNA in a number of studies (Caron *et al.* 2000; Dutech *et al.* 2000; Raspé *et al.* 2000).

Table 3.2 PCR primers showing regions of the chloroplast genome amplified.

Locus	Primer 1	Locus	Primer 2	Sequence length (bp)	Annealing temp. (^{o}C)
tRNA L(UAA)	ttcaagtccctctatccc	tRNA F(GAA)	gtcctctgctctaccaact	~430	56.6
psII 44kd protein	ttccgtgggtggcgtag	tRNA-Ser(UGA)	gaatccctctctctccttt	~450	62.1

3.4 Construction of a Partial Library Enriched for Microsatellites

Libraries enriched for AT/TA and GA/CT repeats of both species were constructed according to Schlötterer (1998). The steps involved are described below.

3.4.1 Digestion of DNA with Restriction Enzymes

Restriction digests were carried out usually in a total volume of 20 µl, using 2-5 U of commercially available enzymes and buffers, and one µl of spermidine. The samples were mixed, centrifuged for a few seconds and incubated at 37 °C for 2-12 h.

For construction of libraries, DNA was required that was 300-700 bp in order to be easily cloned. Therefore, 20 µg of DNA of *H. mantegazzianum* and *I. glandulifera*, which had been extracted as in section 3.2.1, were cut using *MboI* as above, but also including acelated BSA. DNA was run on a 1.2 % TAE gel.

3.4.2 Agarose Gel Electrophoresis

Electrophoresis of DNA samples was performed using 0.8-1.2 % agarose, TAE (40mM Tris-acetate, 1 mM EDTA, pH, 7.6) gels, with TAE as buffer. TBE (Tris-borate, EDTA) was substituted for TAE when DNA recovery from gels was not required). Two drops of a 50 µg/ml solution of ethidium bromide was added to the molten gel to allow DNA to be visualized under UV light (150 mJ, Bio-Rad GS Gene Linker UV chamber). DNA samples were prepared by adding one-sixth volume of loading buffer (30% glycerol, 0.25% bromophenol blue, 0.25% xylene cyanol FF). Gels were run horizontally at 50-100 V. Gels were analysed under UV light and documented using a Bio-Rad Gel Doc 1000 video documentation system.

3.4.3 Excision of DNA from Agarose Gels.

The Geneclean II commercial kit (Anachem) was used in order to obtain DNA from agarose gels. The required band (for libraries, the 300-700 bp band) was cut out of DNA out of an agarose gel, cut into small pieces and placed into 1.5 ml tubes. Three volumes of NaI stock solution were added and tubes were placed into a 45-55 °C water bath for six minutes. Five µl of "glassmilk" was added, mixed and the mixture placed on ice for five minutes. The silica matrix containing bound DNA was pelleted

by centrifugation for five seconds. The supernatant was washed three times with NEW Wash buffer (10-50 volumes). The pellet was resuspended, centrifuged for 5 seconds, washed three times, and incubated in TE at 45-55 °C for 2-3 minutes. Following centrifugation for 30 seconds, the supernatant, containing the DNA, was removed to a new tube.

3.4.4 Preparation of Linkers

Linkers *SauL* A and B were obtained from MWG Biotech, and taken up in a 1μg/μl solution of 0.2 M TE. 5 μg of oligo *SauL* B was phosphorylated using commercially obtained polynucleotide kinase, phosphorylation buffer and 2.5 mM ATP in a 20 μl reaction. The reaction mixture was incubated at 37 °C for 30 minutes. 1μl 0.5M EDTA was added, the DNA precipitated with 0.1 volumes of 3M sodium acetate and four volumes of ethanol, and resuspended in 5μl of water. 5 μg of *SauL* A and 5 μg phosphorylated *SauL* B were incubated in 0.1M NaCl at 60 °C for one hour and ligated immediately afterwards to the restricted DNA of *H. mantegazzianum* and *I. glandulifera.*

3.4.5 DNA Ligation

Ligation of DNA fragments was carried out in 10 μl reactions, using commercially available T4 DNA ligase and buffers (GibCo BRL). Ligations were incubated at 15 °C for 16 hours.

Five μg of linkers were ligated with 500 ng of restricted *H. mantegazzianum* and *I. glandulifera* DNA. The reaction was stopped by heating to 68 °C for 10 minutes. The samples were then immediately cooled.

The ligation mixtures were run on a 1.2 % agarose gel. The 300-700 bp fraction was excised from the gel as previously. The linkers were ligated to the plant DNA in order to amplify the DNA using the linkers as PCR primers.

3.4.6 PCR (Polymerase Chain Reaction)

PCR was performed using standard conditions. Each reaction (10-100 μl) contained reaction buffer containing 1.5mM $MgCl_2$, 10mM Tris-HCl pH 8.4, 50 mM KCl, 0.2mM each of dATP, dCTP, dGTP and dTTP, two primers (each 1μM), template and

DNA Taq polymerase (0.2-1 unit, where unit one catalyses the incorporation of 10 nmol of dNTP into acid-insoluble form in 30 minutes at 74°C). PCR amplifications consisted of an initial denaturing step of 94 °C for five minutes followed by 25-35 cycles of the annealing temperature (48-62 °C) for 1 minute, 72 °C extension temperature for 1-2 minutes and 94 °C for 30 seconds. This was followed by one minute at the annealing temperature, ten minutes at 72 °C and the samples were maintained at 4 °C. Reactions were carried out using an MJ Research thermal cycler.

A PCR amplification, using *SauL* A as a primer, was set up using 35 cycles using a 67 °C annealing temperature and a two minute extension time. Thirty μg of PCR library of each species was obtained, and purified using the Geneclean II kit as previously. The DNA was taken up in 0.2 μg/μl of water in preparation for hybridisation to microsatellite sequences (see below).

3.4.7 Extension of Microsatellite Probes

The microsatellite sequences to be used to probe the libraries were extended in a PCR-like reaction (Schlötterer 1998). Eighteen base pair sequences of GA, CT and AT were used as following a literature search, they were found to be the most common microsatellites in plants. The reaction mixtures were made up to a total volume of 50 μl and contained either 5 ng/μl of GA and CT repeats or 10 ng/μl of AT repeats. In a modification of the procedure described by Schlötterer (1998), 1μM of each primer, 0.2mM of each nucleotide and 2 units of *Taq* were used in each 50 μl reaction. The PCR conditions, with 35 cycles, were as previously described but with a 39 °C annealing temperature. The amplified probes were run on a 1.2 % agarose gel and the 700-2000bp fraction excised and the DNA obtained using Geneclean II. For GA/CT, the yield of this fraction was quite low and so the amplified probe was then itself reamplified.

3.4.8 Hybridisation of PCR library to Amplified Microsatellite Sequences

Nylon filters were used for hybridisation selection in a 100 μl volume. The membranes were cut into pieces.2-3 mm square. 200 ng of each extended microsatellite was used and 0.1 volumes of 1 M KOH were added at room temperature. After five minutes, 0.25 volumes of 1 M Tris-HCl pH 4.8 was added and the samples placed on ice. The solution was spotted, 2 μl at a time, onto the cut pieces

of nylon hybridisation filter (Hybond N or Amersham) using both sides of the filter. Once dried, the DNA was fixed to the filter using UV radiation.

The target filters were prehybridised in a 1.5 ml tube with 1 ml Church buffer (0.5 M Na_2HPO_4, pH 7.0; 7% SDS) at 65 °C for 1-16 hours on a shaker. The buffer was removed and replaced with 100 µl of fresh buffer. To about 1 µg of PCR library in about 5 µl of water, 1 µl of 0.2 M KOH was added. After five minutes at room temperature, the samples were placed on ice and 5 µl of 1 M Tris-HCl pH 4.8 was added. 0.5 µl of 0.25M HCl was then added to the solution containing the filters. Hybridisation was carried out at 65 °C for 16 hours. The hybridisation buffer was removed and the filters washed four times with wash solution (2.5 x SSC, 0.1% SDS). This was followed by four washes in 25 ml for a total of 30 min. After washing, the excess moisture was blotted off and each filter was placed in a tube containing 50 µl of 50mM KOH with 0.01% SDS. After five minutes at room temperature, 50 µl was removed to a fresh tube. Fifty µl of 50mM Tris-HCl pH 7.5 with 0.01% SDS was added and the solution incubated at room temperature for five minutes. The solutions were then pooled. Ten µl of 10µM *SauL* A (an inactive carrier) was added with 10 µl of 2M NaoAc (pH 5.6) and the solution mixed. The enriched DNA was precipitated with 250 µl of ethanol. After centrifugation, the pellet was washed with 80 % ethanol, dried under vacuum and redissolved in 20 µl water.

The DNA was reamplifed using *SauL* A as PCR primers and restricted with *Mbo*I (in order to remove the linkers) as previously. The linkers were separated by running the restriction on a gel and recovering the 300-700 bp band of DNA as explained above. The enriched PCR library could then be cloned onto a cloning vector.

3.4.9 Ligation of PCR Library to Plasmid

Ten µg *Bluescript II* plasmids were cut with *BamH I*, as above and purified using Geneclean II as previously. The ends of the plasmid were dephosphorylated using NEB buffer and CIP (calf intestinal phosphotase) in a total volume of 50 µl. The mixture was incubated at 37 °C for 45 minutes and was then heated to 65 °C for 10 minutes. A phenol/chloroform extraction was carried out and the plasmid taken up in 25 µl TE.

The "ligatemac" program was used to find the optimum ratio of vector to insert. The plasmid and PCR library were ligated at a 6:1 ratio in a total volume of 20

μl. The plasmid and PCR library were first pooled and heated to 45 °C for five minutes, to melt any cohesive termini that had reannealed, and then cooled to 0 °C. T4 ligase buffer and 0.5 μl bacteriophage T4 DNA ligase was then added and the solution incubated at 16 °C overnight. The plasmids were then ready to be inoculated into competent cells.

3.4.10 **Preparation of competent bacterial cells** (Chung *et al.* 1988)

Two ml of XL1-Blue (Stratagene) cells in Luria-Bertani (LB) broth (1% w/v bacteriological peptone, 1% w/v NaCl, 0.5% w/v yeast extract) containing 12.5 μg/ml tetracycline was added to a sterile 250 ml conical flask containing 10ml of LB broth. The mixture was incubated at 37 °C, on a cyclical rotator at 500 cycles/min until OD 600= 0.3-0.6. The cells were pelleted by centrifugation at 1,000 g for 10 minutes at 4 °C. The cells were resuspended in $1/10^{th}$ vol. of TSB (transformation and storage buffer- LB broth pH 6.1, 10% PEG, 5 % DMSO, 20 mM Mg2+ (10 mM MgCl2, 10mM MgSO4)) at 4 °C. The mixture was incubated on ice for 10 minutes.

3.4.11 **Transformation and Expression of Bluescript with Insert.**

Transformation was carried out by pipetting 0.1 ml aliquots of cells into cold polypropylene tubes and mixing with plasmid. The cells were returned to ice for 15-30 minutes. LB agar plates (containing 12.5 μg/ml tetracycline and 60 μg/ml ampicillin) were dried. In order to identify plasmids containing inserts by using blue/white colour selection, 25 μl of a solution containing 20 mg/ml X- gal (5-bromo-4-chloro-3-indolyl β-D-galactose) and 10 μl of a 0.1 mM solution of IPTG (isopropyl-β-D-thiogalactoside) was added to each plate. The mixture of transformed cells, in 110 μl, was spread onto each plate. Plates were incubated for 16 h at 37 °C and then stored at 4 °C .

3.5 Screening of libraries using DIG-Labelled Probes

For library screening with microsatellite probes, the DIG High Prime DNA labeling and detection starter kit II (Roche Cat. No. 1585614) was used. Approximately 1 μg of amplified AT and GA/CT microsatellite repeats was pooled and DIG-labelled, using the DIG High-Prime labeling system to incorporate DIG-11-dUTP according to the supplied protocols. Plates containing transformed *E. coli* were incubated at 37 °C until colonies were visible and then chilled at 4 °C for 30 minutes. A positively charged nylon membrane (Roche) was laid onto the agar and left for one minute, then carefully removed after marking the membrane and dish to allow re-alignment once positive colonies had been identified on the membrane. Throughout the screening protocol, the membrane was placed with the side containing colonies uppermost. The membrane was briefly blotted on dry Whatman 3MM paper, then blotted on 3MM paper soaked with denaturation solution (0.5M NaOH, 1.5M NaCl) for 5 minutes. The membrane was again blotted briefly on dry 3MM paper, then blotted on 3MM paper, soaked in neutralization solution (1.0M Tris-HCl, pH7.5, 1.5 M NaCl) for 15 minutes. Following a further brief blot on dry 3MM paper, the membrane was blotted on 3MM paper soaked with 2 x SSC (0.3 M NaCl, 30mM sodium citrate, pH 7.0) and once again blotted briefly on dry 3MM paper. DNA bound to the membrane was then crosslinked using UV light (150 mJ). To remove any contaminating protein, the membrane was subsequently treated with 0.2 mg/ml proteinase K soution for 1 hour at 37 °C. The treated membrane was blotted between sheets of 3MM paper soaked in distilled water to remove any remaining cellular debris. The membrane was then probed, using a Technel hybridiser HB-1D for all incubations to maintain the correct temperature and agitation. Optimum hybridisation and washing conditions were determined empirically. The membrane was first pre-hybridised at 62 °C for 1 hour using hybridisation buffer consisting of 5x SSC (0.75 M NaCl, 75 mM sodium citrate pH 7.0), 0.1% *N*-lauroylsarcosine, 0.02% SDS, 1% blocking reagent (Roche). The DNA probe in 1 ml hybridisation buffer was then denatured by heating to 100 °C for 10 minutes, and cooled rapidly on ice then added to hybridization buffer pre-warmed to 66°C. The membrane was hybridized in this solution for 16 hours at 62°C. The hybridisation solution was then poured off and stored at –20 °C for re-use, and the membrane was washed to remove weakly bound probe.

Washing consisted of two washes with 2 x SSC containing 0.1% SDS for 5 min at room temperature, followed by a further two washes with 0.5x SSC containing 0.1% SDS for 15 minutes at 45°C. All subsequent steps were performed at room temperature. The membrane was then washed in maleic acid buffer (100mM maleic acid, 150 mM NaCl, pH 7.5 containing 0.3% v/v Tween 20 for 1 minute, then treated with blocking solution (maleic acid buffer containing 1% blocking reagent) for 1 hour. Anti-DIG-alkaline phosphotase antibody was then added to this blocking solution at a dilution of 1:10,000 and the membrane incubated for 30 minutes. The membrane was then washed twice with maleic acid buffer containing 0.3% v/v Triton X for 15 minutes. The membrane was rinsed with development buffer and developed with BCIP and NBT (Nitroblue tetrazolium). The membrane was placed on X-ray film in a development folder for 80-200 minutes.

3.6 Obtaining Microsatellite Sequences

The film was developed and positive colonies were identified by lining up the film with the agar plates on which the colonies were grown. The plasmids were miniprepped as below (Birnboim & Doly 1979).

The positive colonies were placed in 10 ml LB agar broth, incubated at 37 °C for 16 hours and placed at 4 °C overnight. The cells were harvested by centrifugation (at 1,000g, as were all subsequent centrifugations) for 10 minutes at 4 °C. The cell pellet was resuspended in 200 µl of a solution of 50 mM glucose, 25 mM Tris pH 8.0 and 10 mM EDTA pH 8.0. The mixture was transferred to 1.5 ml tubes and left for a few minutes. Four hundred µl of a solution containing 0.2 M NaOH and 1% SDS was added and incubated on ice for 10 minutes. 300 µl of an ice-cold solution of 60 ml 5 M KOAc and 11.5 ml glacial acetic acid was added, the mixture left on ice for 5 minutes and centrifuged for 10 minutes. Eight hundred µl of the supernatant was transferred to a new tube and incubated in a 20 µg/ml solution of RNase A for 2 minutes. Six hundred µl of a solution of phenol/chloroform/isoamyl alcohol (25:24:1) was added. Following centrifugation 10 seconds, the aqueous phase was collected. The extraction was repeated using 600 µl of chloroform after which 480 µl isopropanol was added and the DNA allowed to precipitate for 15 minutes at room temperature. Following centrifugation for 10 minutes, the pellet was washed with 70

% ethanol, dried briefly under vacuum and resuspended in 50 µl of TE. 50 µl of 5M LiCl was added, the mixture placed at –20 °C for 30 minutes and centrifuged for 10 minutes. The supernatant was collected and 200 µl ethanol added. The suspension was incubated at room temperature for 15 minutes and then centrifuged for 15 minutes. The pellet was washed twice in 70% ethanol, centrifuged for two minutes and then resuspended in 30 µl TE.

The insert cut out using *PvuII* restriction enzyme. In order to confirm that these colonies contained microsatellite sequences, the cut plasmid was run on a 1.2 % agarose gel and a Southern blot performed (Sambrook 1989). Probing of the gel with DIG-labelled microsatellite sequences was carried out as above. Inserts confirmed as containing microsatellite sequences were sequenced using an ABI 377A automated fluorescent sequencer.

3.7 Primer Design and Titration

Twenty-three microsatellite loci of *H. mantegazzianum* and 27 of *I. glandulifera* were sequenced. All of the microsatellites of *H. mantegazzianum* consisted of GA/CT repeats, rather than AT repeats. There were three AT repeat sequences of *I. glandulifera*, but in two cases, these sequences were found alongside GA repeats, in a compound microsatellite. The low number of AT repeats found may have occurred because when amplifying the microsatellite sequences, the GA/CT repeats were reamplified following an initially low yield. Therefore, there was a disproportionately large amount of GA/CT repeats relative to AT repeats in the probe mixture and the paucity of resulting AT repeat sequences is unlikely to be a reflection of the relative frequencies of the two types of repeats in the study species.

In order for the microsatellite sequences to be amplified, flanking sequences were required for use as PCR primers. Over 20 microsatellite sequences were obtained for each species. However, many were not useful for a number of reasons:- microsatellite sequences contained too few repeats; one or both of the flanking sequences were too short; or because two compatible PCR primers could not be found (see section 3.3 for requirements of PCR primers). The sequences surrounding microsatellites are often both AT-rich and repetitive in nature, which lead to difficulties in designing primers. As the libraries were amplified after the initial

hybridisation, there were many copies of the same sequences and three of the *H. mantegazzianum* and two of the *I. glandulifera* sequences obtained were found to be repeats.

Primers were designed for seven sequences of each species and PCR carried out as described above. Titration of conditions that were carried out in order to amplify the required band included changing the annealing temperatures and extension times. Three of the loci of *I. glandulifera* and one of the loci of *H. mantegazzianum* could not be amplified, despite considerable efforts altering conditions. The enrichment process may lead to the production of "phantom" microsatellite sequences whereby two different sequences stick together. The resulting sequence does not exist in the plant genome and so cannot be amplified (Koblížková *et al.* 1998). This may be the reason for some of the loci not being able to be amplified.

3.8 Screening for Variation at Microsatellite Loci

Four loci of *I. glandulifera* and six of *H. mantegazzianum* were screened for variation by including FAM-6 labelled dUTPs (ABI, Perkin-Elmer) in the PCR mix. Twelve individuals of each species were amplified with each set of primers. The labelled PCR products were run on a sequencing gel in an ABI 377A automated sequencing machine. The PCR products were sized to the nearest base-pair using the GeneScan™ and Genotyper ™ (ABI) software packages (section 3.9). Variation was found in three loci of *I. glandulifera* and four of *H. mantegazzianum.* All loci contained GA repeats and the loci are shown in Table 3.3.

Table 3.3 Microsatellite Loci. A2, A3 and A21 are loci of *I. glandulifera* and A34, A43, A46 and C52 are those of *H.mantegazzianum.*

Locus	Primer 1	Primer 2	Sequence length (bp)	Annealing temperature (^{O}C)
A2	accacggacgcaagtga	gcaagagaagttggcggaa	332	53
A3	acttccatgtgttattga	tgaaagatgggttacatt	350	50
A21	actcttctggctaagctg	aaagcgagaagttggcg	315	53
A34	tgccttgactattttactt	aaaataaatgataaaatccct	420	50
A43	gacccacaagagaagaagt	gttccaacgaagcctatta	226	50
A46	agctcgggctagtcttc	cactcacaacaatgcagc	285	53
C52	gcaatttctcgacactccc	gcatcatagcgcaactgc	185	56

3.9 Sizing Microsatellite Alleles using Automated Fluorescence

Microsatellite loci from collected leaf samples were amplified by PCR using fluorescently labelled primers. The reverse primer for each locus was labelled with 6-FAM, HEX or NED ABI dyes (Perkin-Elmer) and PCRs were run in a total volume of 10 µl using $1/10^{th}$ labelled and $9/10^{th}$ unlabelled reverse primer using cycling conditions as for section 2.3.6. Using $1/10^{th}$ labelled primer was found to be just as effective as using only labelled primer and the efficiency of labelling reactions was improved.

The labelled PCR products were run on a sequencing gel in an ABI 377A automated sequencing machine. Gels were prepared using 14.4 g urea, 21.1 ml deionised water and 4.5 ml of 40% 29:1 acrylamide, bisacrylamide. One g amberlite resin was added to deionise the solution and was removed using filter sterilisation. In order to polymerise the acrylamide, 4 ml TBE, 200 µl ammonium persulphate and 22.5 µl TEMED were added. The mixture was poured into a 36 cm gel.

The different dyes have different strengths, and so the volume of PCR product run on the gels varied with dyes. PCR products with sizes that do not overlap and products of different colours can be run in the same lane of a gel. 0.4 µl of HEX-labelled product and 0.3 µl of NED and 6-FAM-labelled PCR product were run in each lane. The microsatellite primers were designed and ordered in dyes such that all

seven loci could be run in each lane. Therefore, one individual of each species could be genotyped in each lane.

The gels used were 36 cm wide and were run at 41°C for 3 hours on an ABI 377A automated sequencing machine using Data Collection software and the virtual matrix, filter set C or D. Each lane was loaded with one PCR product from each locus and the internal size standard TAMRA-500 (for use with filter set C) and ROX-500 (for use with filter set D). The ABI 377A machine uses a laser to scan products running off the gel at regular intervals. The labelled PCR-products show up as peaks of different colours, depending upon the fluorescent dye used.

Data from the gels were analysed using GeneScan™ Analysis software (ABI) The filter set used was first specified, in order to ensure that different dyes appear as different colours that do not leak into one another. However, there is always some overlap and so if too much of one dye is run on a lane, it may also show up as another colour. When using filter set D, 6-FAM appears blue, HEX appears green, NED is yellow and ROX is red. The lanes of the gel were then tracked automatically. The speed at which products run down a gel is proportional to their size. The ROX-500 size standard contains products with peaks at 35, 50, 100, 139, 150, 160, 200, 250, 300, 340, 350, 400, 450 and 500 bp. The data from the laser is collected in terms of the number of scans that had been made when peaks are detected.

The internal size standards are used to convert number of scans into base pairs within each lane in order to allow for lane to lane variation which can occur when gels do not run straight. Using the points of known size in each lane, the software creates a calibration curve that enables the size of any product detected along the gel to be determined to the nearest base pair (Figure 3.3). The data showing the sizes of the different peaks was then transferred to the ABI software package, Genotyper™. This enables all lanes run to be viewed together, for each dye colour used and shows the size of each peak in base pairs. An example of a Genotyper™ file is shown in Figure 3.4, which shows the alleles of locus A3 of four individuals of *I. glandulifera*. The sizes of alleles are given in boxes and can also be read using the scale bar. The alleles sizes are not integers and there can be variations between gels of up to 0.6 base pairs. Therefore, allele sizes were put into allele class categories, eg. the alleles in Figure 3.4 were assigned sizes of 342, 348 and 350 base pairs.

Figure 3.3 GeneScan Gel Showing the amplified nuclear microsatellite loci of *H. mantegazzianum* (A34, A43, A46, C52) and *I. glandulifera* (A2, A3, A21). The loci were amplified using fluorescently labelled PCR primers, which enable them to be detected by laser. The red bands are size standards, which are used to produce a calibration curve enabling all bands to be sized accurately.

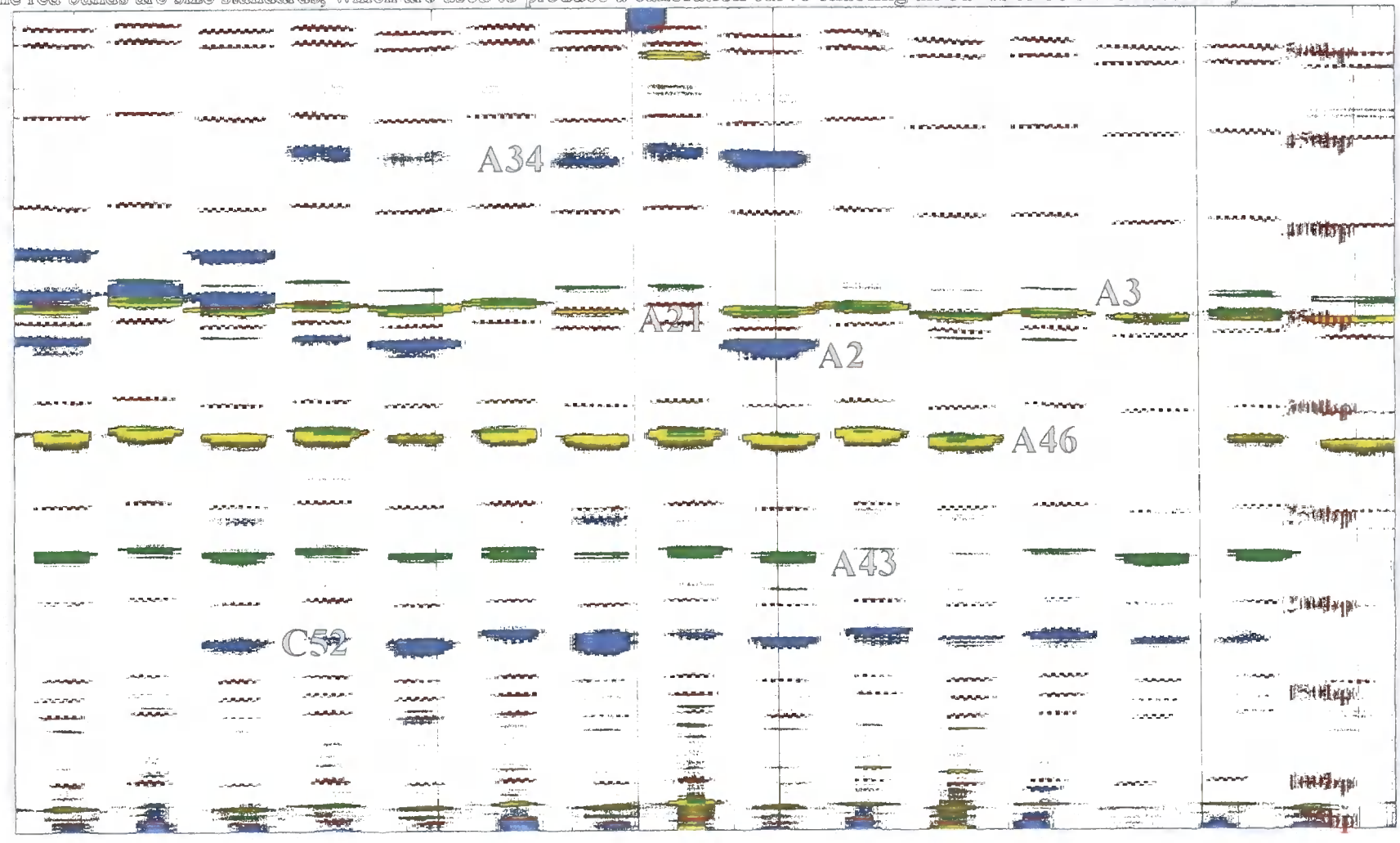
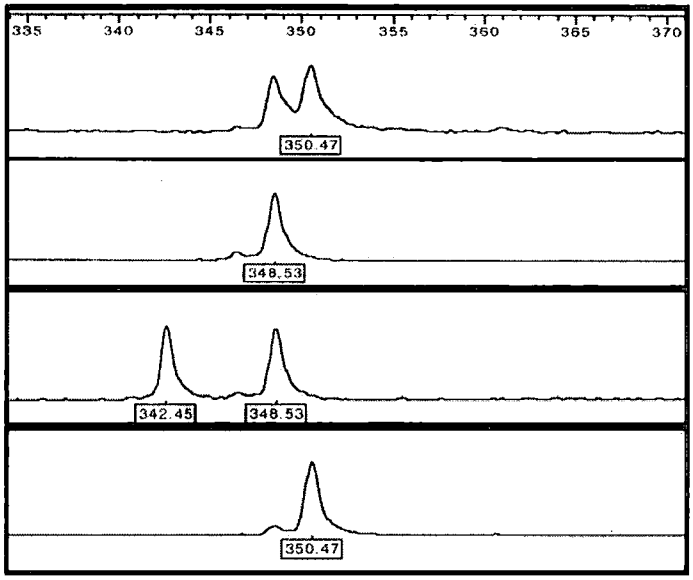


Figure 3.4 Genotyper™ file showing alleles of *I.glandulifera* of the microsatellite locus A3. The numbers along the top and in the boxes are sizes in base pairs.



3.10 Chloroplast Microsatellite Loci

Weising and Gardner (1999) designed universal primers to ten microsatellite loci in plants. Variation at three of these loci in *H. mantegazzianum* and *I. glandulifera* was tested for by amplifying these loci using labelled dUTPs and sizing the PCR products using GeneScan™ as described in section 3.9. Variation was found at one locus of each species. The reverse primers of these loci were labelled with ABI dyes and individuals screened for variation alongside the other microsatellites. One chloroplast microsatellite was found to be polymorphic for each species, and reverse primers were ordered which were labelled with ABI dyes. The chloroplast microsatellite loci were smaller than the other loci and so were able to be run in the same lane as the other loci. Forty μl of water was added to the 10 μl reaction tubes following PCR, and 0.3 μl were run on the ABI gels.

3.11 Genetic Analysis

3.11.1 Measurement of Genetic Variation

There are a number of statistical tests that can be used to analyse genetic variation at the population level. The amount of differentiation between populations can be quantified in terms of F_{ST}, which does not consider different mutational relationships among alleles (section 2.3). The program Genepop version 3.2 (Raymond & Rousset 1995) was used to calculate values of F_{ST}.

An analogous measure that also takes into account relative allele size is R_{ST} (Slatkin 1995), which was defined in section 2.3. However, this measure relies on the assumptions that all populations are the same size and that all loci have equal variances. Since these assumptions may be violated in this study, an unbiased estimator *Rho* (Rousset 1996) was used, which does not rely on the above assumptions. Differences in population size are controlled for and the disparity in variance between loci is accounted for by transforming the data to express alleles in terms of standard deviation from a globalised mean (Goodman 1997). Another measure of genetic differentation is $(\delta\mu)^2$ which is the sum of the square of the difference between the average repeat size in two populations populations (section 2.3) (Goldstein *et al.* 1995b). Values of $(\delta\mu)^2$ and Rho_{ST} were calculated using the RST Calc. program (Goodman 1997).

Many statistical tests used for genetic analysis rely on the assumption that levels of gene flow and genetic drift are at equilibrium. If this is the case then the rate of gene flow will be inversely related to the differentiation between populations, and so gene flow can be estimated using the relationship

$$F_{ST} = 1/(1 + 4Nm) \qquad (1)$$

where N is the effective population size and m the proportion of immigrants in a population so that Nm is the number of migrants (Wright 1978). The Arlequin version 2.000 (Schneider *et al.* 2000) and RST Calc. (Goodman 1997) packages calculate values of Nm.

Considering the relatively short time that *H. mantegazzianum* has been present in the Northeast of England and that populations may be ephemeral, there is unlikely to be an equilibrium between migration and genetic drift, which is an assumption of all the tests described above. Assignment tests are not based on such assumptions, and

are based on multilocus data so are well suited for analysing microsatellite variation (Paetkau *et al.* 1995; Davies *et al.* 1999). Assignment tests calculate the likelihood of drawing a particular multilocus genotype from a number of potential populations. The more loci included, the more accurate the test is likely to be (Boeklen & Howard 1997). The results of pairwise comparisons can be shown graphically by plotting the log-likelihood of each genotype in each population being drawn from its own population versus the other population. This method can be used to identify individuals in populations whose genotypes suggest they are immigrants. In the case of populations of *H. mantegazzianum* along rivers, individuals that have dispersed from populations upstream could potentially be identified. The tests can also provide a measure of differentiation between populations. Isolation by distance can be looked at by calculating a regression of $F_{ST}/(1- F_{ST})$ estimates to either geographic distances or its natural logarithm. Arlequin version 2.000 (Schneider *et al.* 2000) was used to perform assignment tests and tests of isolation by distance.

3.11.2 Tests of Hardy-Weinberg Equilibrium

Interpretations of patterns of genetic variation based on the Infinite Island Model assume that populations are in Hardy-Weinberg equilibrium. There are a number of possible causes of deviations from the Hardy-Weinberg equilibrium (section 2.1.3). Exact tests of Hardy-Weinberg equilibrium (Guo and Thompson 1992) using a Markov chain with a forecasted chain length of 100,000 and 1,000 dememorization steps were carried out using Arlequin version 2.000.

3.11.3 Detection of Bottlenecks

Luikart *et al.* (1998) presented a graphical method for detecting recent bottlenecks in populations. The frequencies of each allele of multilocus genotypes in each population were calculated. These frequencies were converted to proportions and these were placed into size classes. The class sizes were:-

0-0.1; 01 1-0.2; 0.21-0.3; 0.31-0.4; 0.41-0.5; 0.51-0.6; 0.61-0.7; 0.71-0.8; 0.81-0.9; 0.91-1.

An example of this process would be if, in a population of 30 individuals, one multilocus allele was found three times. This allele would have a frequency of 3/60

(as each individual has two alleles) and so its proportion in the population would be 0.05. It would therefore be placed in the 0-0.1 size class. The number of alleles in each size class was then plotted on a bar chart (Figures 4.11).

Luikart *et al.* (1998) showed that populations which had recently been through a bottleneck had a greater number of alleles in one of the intermediate frequency classes than in the lowest frequency class. This occurs because, following a bottleneck, many of the rarest alleles disappear and other previously rare alleles become more common, so there is a skew in the pattern of allele frequencies. Therefore, where the size class 0-0.1 was not the highest bar (excluding the 0.9-1 size-class bar), the population has been through a bottleneck.

3.11.4 Linkage Disequilibrium

Tests for linkage disequilibrium between all pairs of loci for each species were carried out using Genepop 3.2. A probability test was performed on all pairs of loci using a Markov chain, which gives rise to a *P*-value estimate.

3.12 Genetic Distance Trees

Trees showing genetic distance were constructed using a number of different estimates of genetic distance using PHYLIP (Phylogeny Inference Package) version 3.5c (Felsenstein 1993). The programs in PHYLIP enable the construction of trees based on both Nei's genetic distance estimates and maximum likelihood methods (section 2.3.5). The raw data required for these programs were the numbers and proportions of all (genomic) microsatellite alleles present in each population.

When constructing trees with large numbers of populations, the possible number of tree topologies is so large that it is not possible to test all of them. The programs in the PHYLIP package use the following procedure to find the best tree. Three populations are considered initially and a tree constructed containing only those. The fourth population is then taken and tested at all possible positions and the one which best fits is chosen. The fifth population is then added in the same manner. When all subseqent populations are added, all possible local rearrangement are tested in order to improve the tree. Wherever a new arrangement is an improvement, it is retained.

3.12.1 Construction of Trees based on maximum likelihood.

The CONTML program estimates genetic distance by the restricted maximum likelihood method (section 2.3.5). The units of length are amounts of expected accumulated variance. The lengths of the branches of the trees are calculated by considering only the curvature of the likelihood surface as the length of the branch is varied, holding all other lengths constant, and so may underestimate the variance. The trees produced are unrooted. The trees produced using this program can be redrawn using the DRAWTREE program, which plots unrooted trees whilst allowing the orientation of the tree to be varied. The rotation of the tree and the angle through which the tree is plotted can be chosen

3.12.2 Construction of Trees based on Pairwise Genetic Distances

The GENDIST program was use to compute genetic distance between populations and the results were used to produce trees using the FITCH, KITCH and NEIGHBOR programs.

3.12.3 GENDIST

The GENDIST program computes genetic distances using Nei's genetic distance, which measures the accumulated differences in alleles at each locus between a pair of populations (Nei 1972). The differences between populations are assumed to arise from genetic drift. Nei's distance is formulated for an infinite alleles model of mutation, in which there is a rate of neutral mutation and each mutant is to a completely new allele. The result is a distance matrix comparing populations, which was produced for each species. These matrices were fed into three different programs for producing trees.

The FITCH program estimates distances using an "additive tree model" (section 2.3.5). KITSCH uses the same method as FITCH but also assumes an evolutionary clock, an "ultrametric" model. NEIGHBOR uses a neighbour joining method without the assumption of a clock, using the Neighbor-Joining method version 3.573c. All programs produce unrooted trees.

Chapter Four

Results I: *Heracleum mantegazzianum*

4.1 Populations, Catchments and Microsatellite Loci

Thirteen populations of *H. mantegazzianum* were genotyped using the four polymorphic microsatellites identified from a genomic library and one chloroplast microsatellite (Chapter 3). There was no variation observed in either of the non-microsatellite cpDNA markers (secton 3.3). The total number of individuals sampled was 372. The populations and their locations are shown in Figure 4.1. One population was sampled from north London. The individuals were taken from a large cluster covering over one square mile of wasteland in Perivale, north London. The species was recorded as having been brought into a stately home in the area in 1937, which predates the first recorded introduction of the species into the Tees and Wear catchments. The names of the populations refer to the section on the rivers in which they occur. The rivers were divided up into 500 m long sections by an Environment Agency survey which was used to identify the locations of the species. Population Tees 162 was sampled in July 1997 and resampled as population Tees 162a in April 1999.

The frequencies of alleles of the loci for each population are given in Table 4.1. The names of the loci are A34, A43, A46, C52 and C10. C10 is a mononucleotide chloroplast microsatellite locus. The number of alleles at each locus for each population is shown in Figure 4.2. The average number of alleles per locus per population was 3.98.

Figure 4.1 Distribution of populations of *H. mantegazzianum* in the study area

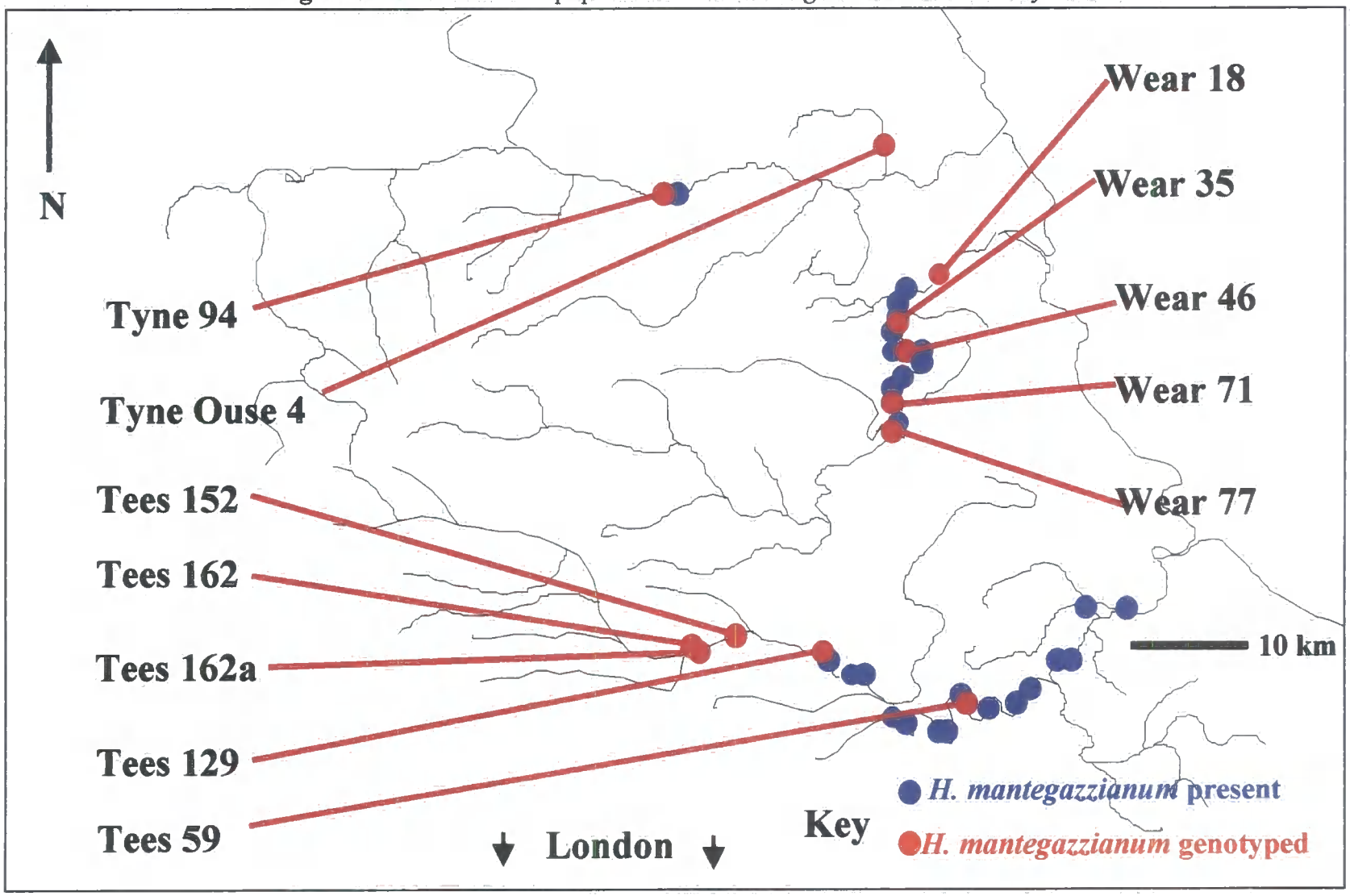


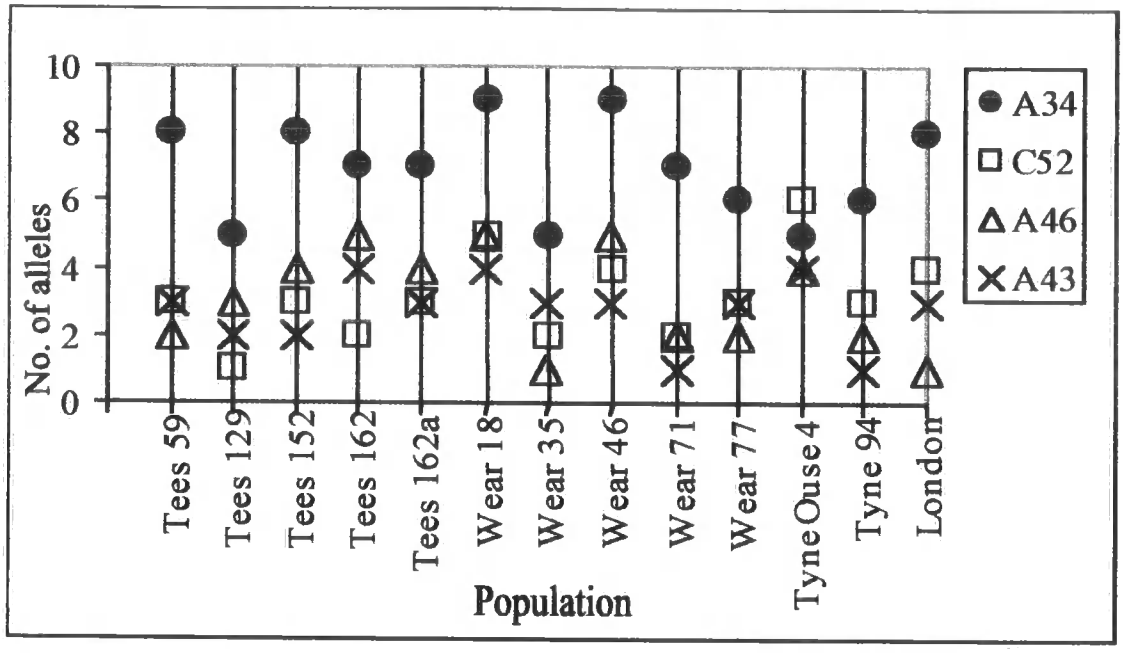
Table 4.1 Table of Allele Frequencies found in each population at each locus. Allele sizes are numbers of base pairs of the locus. Populations are divided into catchments. See section 4.4 for population codes.

Locus		TS59	T129	T152	T162	T162a		WR18	WR35	WR46	WR71	WR77	TOU4	TY94	LND	
A34	**354**			3	16	13	11							7		**354**
	356				1									34	1	**356**
	370								17							**370**
	376							4						1	5	**376**
	378							8	22	4	22			15	2	**378**
	380							1								**380**
	408	1														**408**
	412													1		**412**
	418	1				1				1					6	**418**
	420	7						1							32	**420**
	422	10						2	2	3	1				5	**422**
	424	2			1			19			1	3		1	8	**424**
	426	7			1		2	7	4	7	4	8	1	4		**426**
	428	25				3		8		16	25	16		15		**428**
	430	3	4	8	9	8		8		10	3	16		14		**430**
	432		11	14	12	22				3	4	13		16	1	**432**
	434		16	10	7	12			15	3		4		2		**434**
	436		4	8	13	6										**436**
	438		6	1	2	4										**438**
	440						1			1						**440**

Locus		TS59	T129	T152	T162	T162a	WR18	WR35	WR46	WR71	WR77	TOU4	TY94	LND	
C52	**134**						13		2						**134**
	170	8		1			9	27	14	6	19	1		6	**170**
	172										3				**172**
	174													1	**174**
	178											2			**178**
	180	2				1	32	33	27	58	38	17	12	51	**180**
	182	46	44	44	49	48	2		5			20	28		**182**
	184			15	11	17						18	12		**184**
	186						2					2		2	**186**
A46	**255**		3	5	5	6									**255**
	261				1										**261**
	263						1								**263**
	265	1							4			2	2		**265**
	267											1			**267**
	281						2								**281**
	283			1	1	1	2		7			6			**283**
	285	55	40	49	42	52	52	60	34	59	58	51	50	60	**285**
	287		1	5	11	7	1		1	1	2				**287**
	289								2						**289**

Locus		TS59	T129	T152	T162	T162a	WR18	WR35	WR46	WR71	WR77	TOU4	TY94	LND	
A43	**196**	2													**196**
	198		22	24	26	25		8	1			1			**198**
	200				1										**200**
	222											4			**222**
	224				1										**224**
	226	52	22	36	32	39	51	51	46	60	49	41	52	48	**226**
	228					1	1				7			10	**228**
	230	2													**230**
	232								1		4			2	**232**
	236							1				14			**236**
	252						2								**252**
	254						4								**254**
C10	**97**											28			**97**
	98	28	22	30	30	33	29	30	24	30	30		26	30	**98**
	99											1			**99**
	100											1			**100**

Figure 4.2 Chart showing the number of alleles at each nuclear microsatellite locus in each population.



4.2 Linkage Disequilibrium

The statistical tests carried out on microsatellite variation assume that all loci are independent of one another. Genepop 3.2 was used to test for linkage disequilibrium between the microsatellites. The program tests the null hypothesis that all loci are independent of one another. As none of the *P*-values were significant, the loci can be assumed to be independent.

Table 4.2 Probability test for linkage disequilibrium for each locus pair across all populations (Fisher's method).

Locus pair	Chi^2	df	*P*-value
A34 & A43	22.5	22	0.433
A34 & A46	18.1	22	0.698
A43 & A46	10.7	18	0.908
A34 & C52	33.0	24	0.104
A43 & C52	18.0	20	0.586
A46 & C52	24.6	20	0.219

4.3 Catchment Analysis

In order to compare genetic variation between catchments, all populations in each catchment were pooled.

4.3.1 F_{ST} values of catchment comparisons

F_{ST} values were computed using the Arlequin program, the results of which can be seen in the table below.

Table 4.3 Matrix of F_{ST} values of catchment comparisons

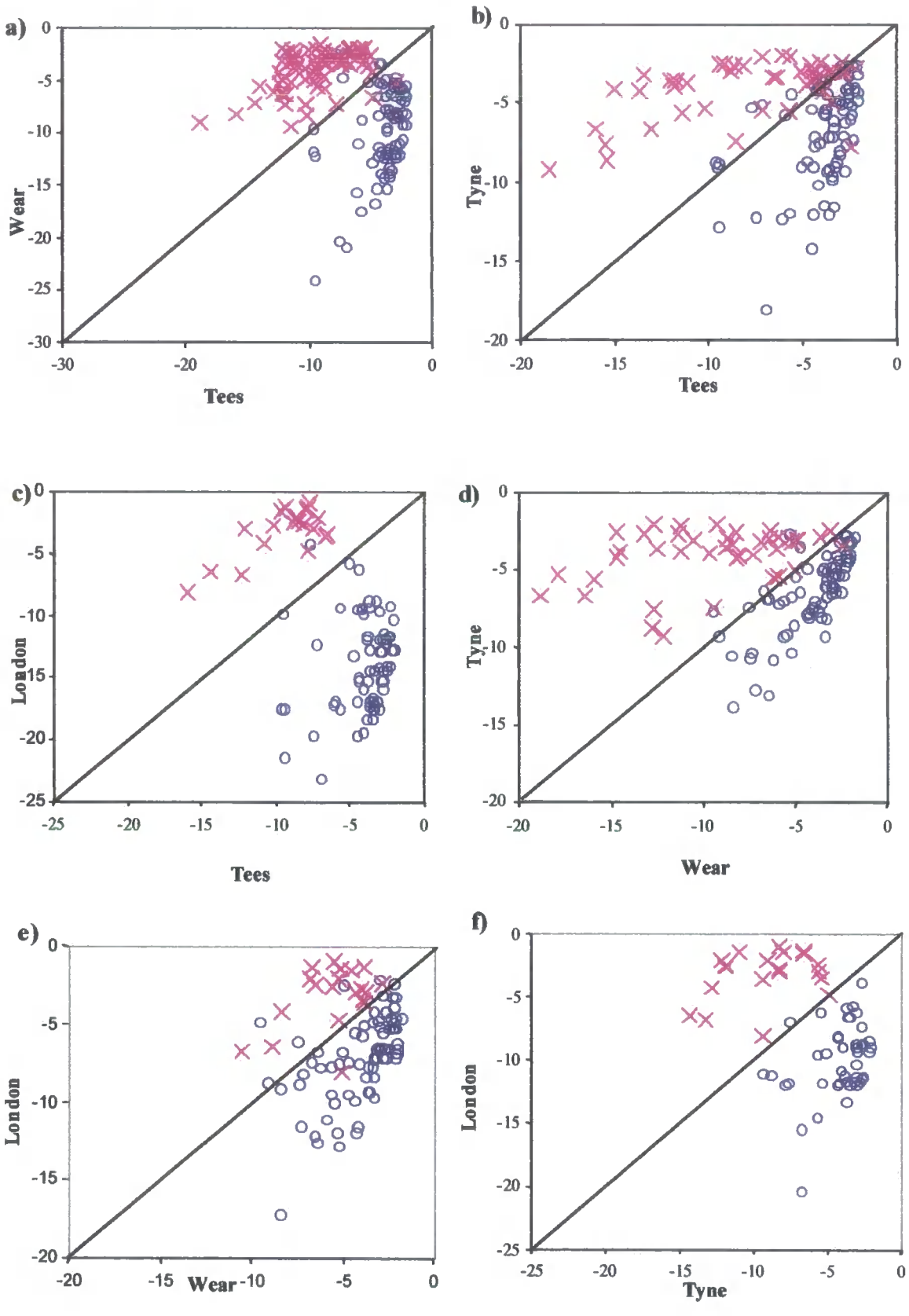
	Tees	Wear	Tyne
Wear	0.27		
Tyne	0.12	0.14	
London	0.35	0.13	0.24

There were significant differences between all catchments. Populations from the Wear were found to be more similar to the population from London than to populations from the Tees.

4.3.2 Assignment tests

Assignment tests were carried out by plotting the log likelihood of genotypes from two catchments. Any genotypes that cross the diagonal line from other genotypes from their own catchment are more similar to the other catchment than to their own. The differentiation of genotypes between catchments can be seen in pairwise plots of maximum likelihood (Figure 4.3). Although there is some overlap in all comparisons, the separation between catchments is surprisingly evident. A comparison of graphs a) and d) shows that the Wear is more similar to London than it is to the Tees, which was also borne out by analysis of F_{ST} values.

Figure 4. 3 Genotype Assignment Tests. Maximum likelihood analysis of individual genotypes from each pair of catchments originating from their own catchment. Comparisons shown are a) Tees v Wear, b) Tees v Tyne, c) Tees v London, d) Wear v Tyne, e) Wear v London, f) Tyne v London. Units are log likelihood values of individuas belonging to each population. Any genotypes that cross the diagonal line from other genotypes from the same catchment are more similar to the other catchment than to their own.

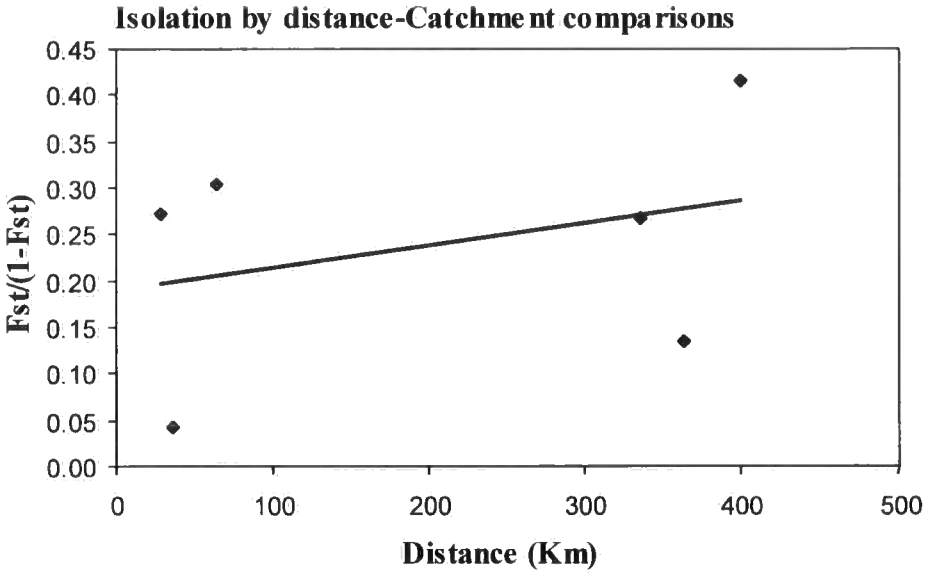


4.3.3 Isolation by distance

The distance between catchments was calculated by calculating the midpoint of the distribution of the species in the Tees, Tyne and Wear river catchments. The distance between these were measured. In the case of the population from London, the distance to the other catchments were measured from the midpoints to Perivale, where the samples were collected.

Plots of F_{ST} /(1- F_{ST}) against distance (Figure 4.4) show no significant relationship between geographic and genetic distance. The population from London was much more distant from any of the other populations, but the genetic distance between the London population and all others did not reflect this.

Figure 4.4 Isolation by distance analysis of catchments. Points are the results of pairwise comparisons between catchments. The equation for the best fit line is y=0.0002x + 0.1892, R^2= 0.112, *P*=0.519.



4.4 Population analysis

In the following figures and tables, the populations were coded as follows:

Tees 59-	TS59	Wear 18-	WR18	Tyne Ouse 4-	TOU4
Tees 129-	T129	Wear 35-	WR35	Tyne 94-	TY94
Tees 152-	T152	Wear 46 -	WR46	London-	LND
Tees 162-	T162	Wear 71-	WR71		
Tees 162a-	T162a	Wear 77-	WR77		

4.4.1 Pie Charts of allele frequencies

Allele frequencies of four loci for populations Tees 129, 152 and 162 are shown in Figure 4.5. Populations Tees 152 and 162 can be seen to be very similar for all loci. Population Tees 129 is also quite similar to the other two populations, but some differences are apparent. The three populations are 16.5 km apart, but are near the most upstream point of the distribution of *H. mantegazzianum* in the Tees.

Locus A34 was the most variable and allele frequencies for nine populations bore this out because there were large differences between many of the populations (Figure 4.6). In population Tyne Ouse 4, the most common allele is found nowhere else. Population Wear 46 contains alleles that appear to be a mixture of those found in the two populations upstream of it which are populations Wear 71 and 77. The most common allele in population Wear 35 was not found in any other population.

Figure 4.5 Alleles (coloured segments) of four microsatellite loci of *H. mantegazzianum* in three populations in the Upper Tees. The white pie segments represent all alleles found at a frequency of 5 % or less.

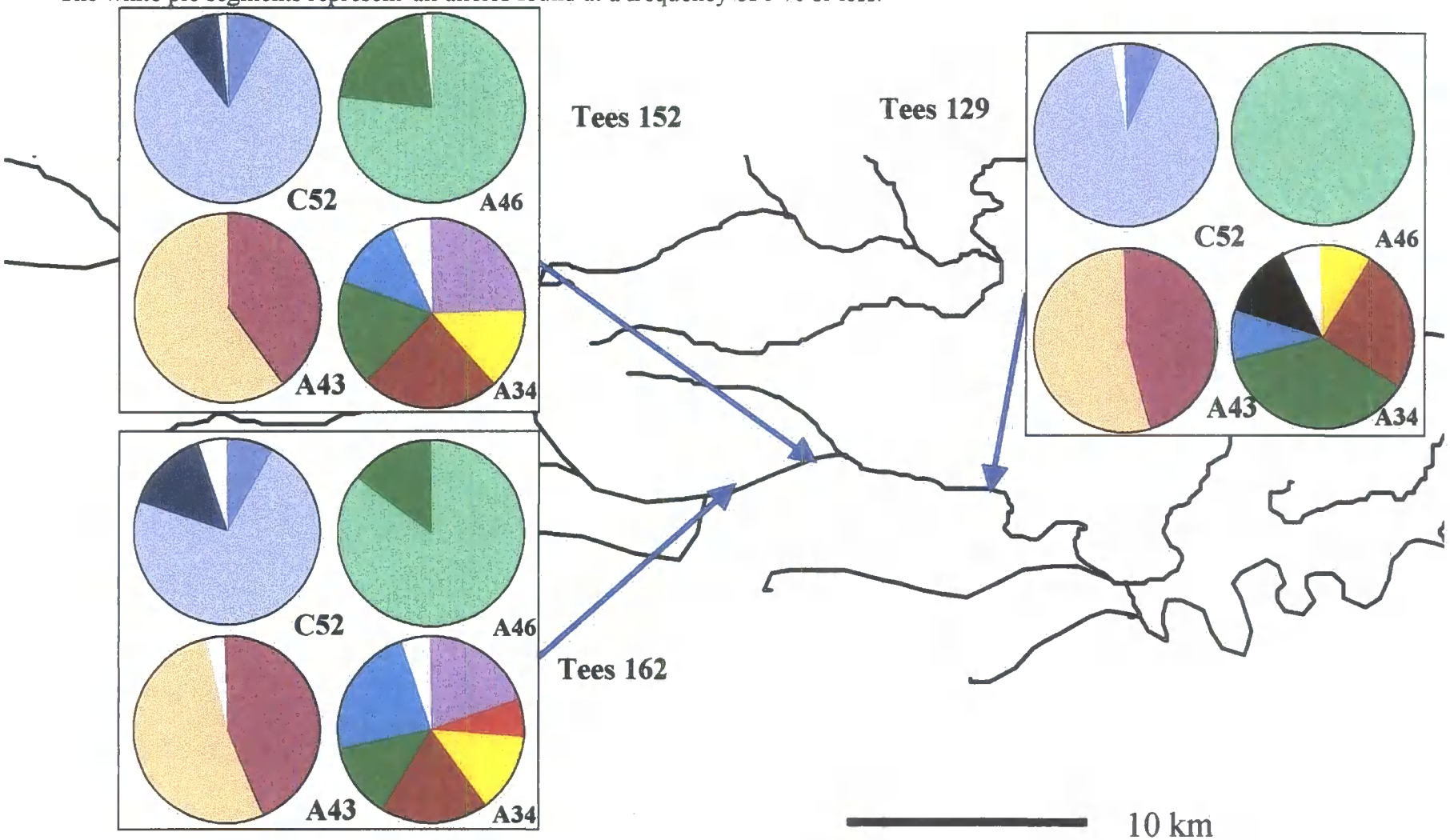
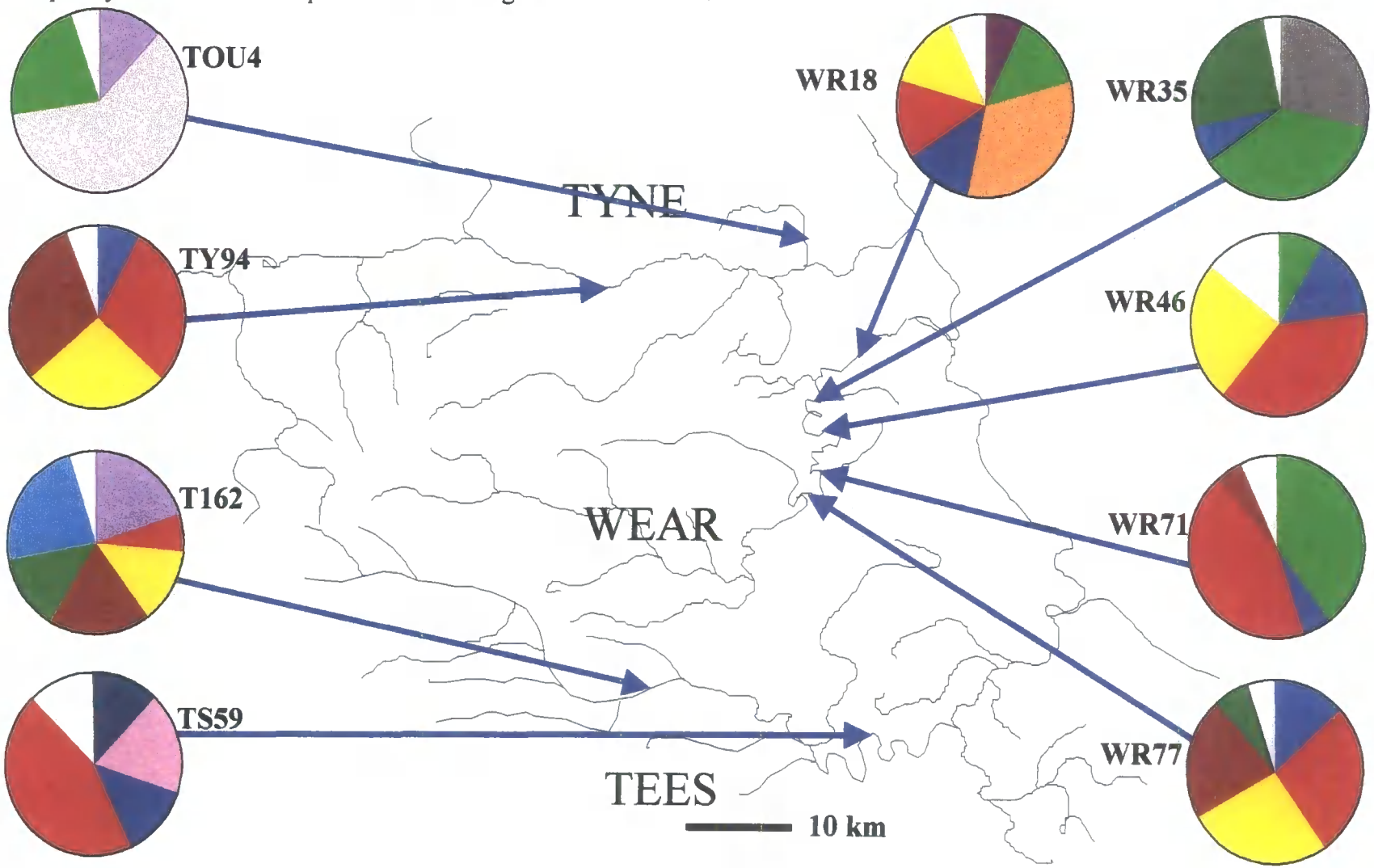


Figure 4.6 Alleles (coloured pie segments)of locus A34 in nine populations. The white pie segments represent all alleles found at a frequency of 5 % or less. Population codes are given in section 4.4.



Population Tees 59 was markedly different from Tees 162 as the second, third and fourth most common alleles found in Tees 59 were not present in Tees 162 and it's most common allele was only the sixth most common in Tees 162.

4.4.2 F_{ST} values of population comparisons

Table 4.4 Matrix of F_{ST} values. Values in bold are within-catchment comparisons. Values in italics are non-significant differences at the level of $P<0.05$. Number of permutations=1000.

	TS59	T129	T152	T162	T162a	WR18	WR35	WR46	WR71	WR77	TOU4	TY94
T129	**0.26**											
T152	**0.19**	**0.05**										
T162	**0.20**	**0.05**	***0.00***									
T162a	**0.19**	**0.04**	***-0.01***	***0.00***								
WR18	0.27	0.38	0.27	0.29	0.27							
WR35	0.36	0.41	0.31	0.34	0.31	**0.12**						
WR46	0.23	0.36	0.25	0.27	0.25	**0.06**	**0.13**					
WR71	0.41	0.53	0.41	0.42	0.40	**0.12**	**0.19**	**0.10**				
WR77	0.29	0.38	0.27	0.29	0.26	**0.06**	**0.12**	***0.03***	**0.12**			
TOU4	0.26	0.29	0.18	0.21	0.19	0.18	0.22	0.19	0.28	0.21		
TY94	0.12	0.26	0.15	0.17	0.13	0.18	0.28	0.14	0.29	0.16	**0.19**	
LND	0.41	0.49	0.38	0.40	0.38	0.14	0.22	0.19	0.22	0.16	0.28	0.33

Apart from population Tees 59, all populations in the Tees and the Wear were more similar to other populations in their catchment than to any other population (Table 4.4). All population comparisons that were not significantly different were between populations in the same catchment. The two populations in the Tyne showed no particular similarity. However, population Tyne Ouse 4 is in a tribuary of the Tyne and is very far from populations on the main river and so is very likely to have resulted from an independent introduction. As was found with catchment comparisons, pairwise comparisons with the population from London was not found to be notably different from those between other populations from different catchments. F_{ST} values for comparisons between populations from the Tees and the Wear were notably large (0.23-0.53).

4.4.3 Rho_{ST} and $(\delta\mu)^2$ Population Comparisons

Rho ST and $(\delta\mu)^2$ values are given in the matrices below using the software program RST Calc. (section 3.11.1). As in section 4.4.2, within catchment comparisons are shown in bold.

Table 4.5 Matrix of Rho_{ST} comparisons.

	TS59	T129	T152	T162	T162a	WR18	WR35	WR46	WR71	WR77	TOU4	TY94
T129	**0.44**											
T152	**0.33**	**0.00**										
T162	**0.34**	**0.00**	**0.00**									
T162a	**0.31**	**0.00**	**0.00**	**0.00**								
WR18	0.19	0.35	0.26	0.28	0.25							
WR35	0.48	0.34	0.24	0.27	0.26	**0.21**						
WR46	0.09	0.39	0.28	0.30	0.26	**0.03**	**0.32**					
WR71	0.35	0.43	0.30	0.33	0.30	**0.10**	**0.13**	**0.15**				
WR77	0.11	0.51	0.38	0.40	0.36	**0.11**	**0.46**	**0.01**	**0.31**			
TOU4	0.71	0.52	0.39	0.41	0.40	0.36	0.04	0.53	0.19	0.76		
TY94	0.08	0.49	0.36	0.37	0.34	0.17	0.45	0.07	0.29	0.25	**0.74**	
LND	0.00	0.44	0.33	0.35	0.31	0.12	0.42	0.02	0.25	0.00	0.62	0.00

Populations in the Tees, with the exception of population Tees 59, are found again to be very similar, but populations in the Wear do not appear to be more similar to one another than to populations in other catchments.

Table 4.6 Matrix of $(\delta\mu)^2$ values. Within catchment comparisons are shown in bold.

	TS59	T129	T152	T162	T162a	WR18	WR35	WR46	WR71	WR77	TOU4	TY94
T129	**47**											
T152	**83**	**83**										
T162	**59**	**41**	**7**									
T162a	**35**	**21**	**22**	**5**								
WR18	64	163	105	111	105							
WR35	243	328	92	147	189	**130**						
WR46	7	61	89	68	45	**37**	**226**					
WR71	60	141	37	55	62	**39**	**65**	**59**				
WR77	8	60	133	99	66	**69**	**309**	**8**	**98**			
TOU4	944	1107	586	725	826	705	252	927	533	1084		
TY94	6	50	126	90	56	99	325	17	103	6	**1091**	
LND	29	104	43	48	44	31	109	28	6	56	648	59

There is even less apparent spatial structure to the results found by analysis of $(\delta\mu)^2$, as population Tees 129 is more similar to Wear 46 than to it's nearest neighbour, Tees 152.

4.4.4 Test of Hardy-Weinberg equilibrium

Table 4.7 shows the result of an exact test of Hardy-Weinberg equilibrium using a Markov chain with a forecasted chain length of 100,000 and 1,000 dememorization steps (Arlequin 2.000). A Bonferroni adjustment for type I errors was carried out to find significant deviations from the Hardy-Weinberg equilibrium (Rice 1989). All loci had significant deviations showing a deficiency in heterozygosity in at least one population. Estimates of Fis were produced using Genepop version 3.2.

Table 4.7 Test of conformity of heterozygosity levels to the Hardy-Weinberg equilibrium. Values in bold are significant deviations at α=0.05 after Bonferroni correction for Type I errors. #Indiv= Number of individuals; Obs. Heter= observed heterozygosity; Exp. Heter= expected heterozygosity.

	Locus	#Indiv	Est. Fis	Obs.Heter.	Exp.Heter.	*P* value	s.d.
Tees 59	A34	28	0.237	0.57	0.75	0.0398	0.0000
	A43	28	1	0	0.17	**0.0002**	0.0000
	A46	28	-0.061	0.04	0.07	1.0000	0.0000
	C52	28	0.077	0.29	0.34	0.0338	0.0004
Tees 129	A34	22	0.425	0.45	0.84	**0.0000**	0.0000
	A43	22	0.114	0.45	0.56	0.6839	0.0015
	A46	22	-0.057	0.18	0.21	1.0000	0.0000
	C52	Monomorphic- no test done					
Tees 152	A34	30	0.397	0.5	0.82	**0.0000**	0.0000
	A43	30	0.045	0.47	0.52	1.0000	0.0000
	A46	30	0.076	0.3	0.35	0.4881	0.0016
	C52	30	0.757	0.1	0.44	**0.0000**	0.0000
Tees 162	A34	30	0.249	0.63	0.85	**0.0000**	0.0000
	A43	30	0.005	0.53	0.54	0.4985	0.0014
	A46	30	-0.049	0.5	0.53	1.0000	0.0000
	C52	30	0.892	0.03	0.36	**0.0000**	0.0000
Tees 162a	A34	33	0.097	0.76	0.81	0.0659	0.0006
	A43	33	0.167	0.45	0.52	0.8262	0.0011
	A46	33	-0.178	0.42	0.37	0.8233	0.0012
	C52	33	0.702	0.12	0.41	**0.0000**	0.0000
Wear 18	A34	29	0.380	0.52	0.83	**0.0000**	0.0000
	A43	29	0.849	0.03	0.26	**0.0000**	0.0000
	A46	29	-0.053	0.21	0.2	1.0000	0.0000
	C52	29	0.403	0.38	0.66	**0.0000**	0.0000
Wear 35	A34	30	0.312	0.6	0.75	**0.0005**	0.0001
	A43	30	0.247	0.17	0.35	0.0960	0.0009
	A46	Ths locus is monomorphic- no test done					
	C52	30	0.214	0.43	0.5	0.4798	0.0016
Wear 46	A34	24	0.479	0.5	0.84	**0.0000**	0.0000
	A43	24	-0.028	0.08	0.08	1.0000	0.0000
	A46	24	0.436	0.33	0.5	0.0328	0.0005
	C52	24	0.205	0.54	0.63	0.0171	0.0004

Population	Locus	#Genot	Est. Fis	Obs.Heter.	Exp.Heter.	*P* value	s.d.
Wear 71	A34	30	0.088	0.73	0.69	0.5718	0.0013
	A43	Ths locus is monomorphic- no test done					
	A46	30	0.429	0.07	0.19	**0.0003**	0.0001
	C52	30	-0.038	0.2	0.21	1.0000	0.0000
Wear 77	A34	30	0.337	0.53	0.8	**0.0004**	0.0001
	A43	30	0.199	0.3	0.37	0.1907	0.0009
	A46	30	-0.067	0.2	0.22	1.0000	0.0000
	C52	30	0.476	0.27	0.52	0.0040	0.0002
Tyne Ouse 4	A34	30	0.240	0.47	0.62	**0.0000**	0.0000
	A43	30	0.17	0.4	0.48	0.5966	0.0015
	A46	30	0.140	0.23	0.33	0.5164	0.0014
	C52	30	0.455	0.4	0.77	**0.0002**	0.0000
Tyne 94	A34	26	0.547	0.35	0.76	**0.0000**	0.0000
	A43	This locus is monomorphic- no test done					
	A46	26	-0.020	0.08	0.11	1.0000	0.0000
	C52	26	0.512	0.31	0.65	0.0014	0.0001
London	A34	30	0.370	0.43	0.72	**0.0004**	0.0001
	A43	30	0.11	0.27	0.34	0.0221	0.0005
	A46	This locus is monomorphic- no test done					
	C52	30	0.140	0.23	0.33	0.0539	0.0008

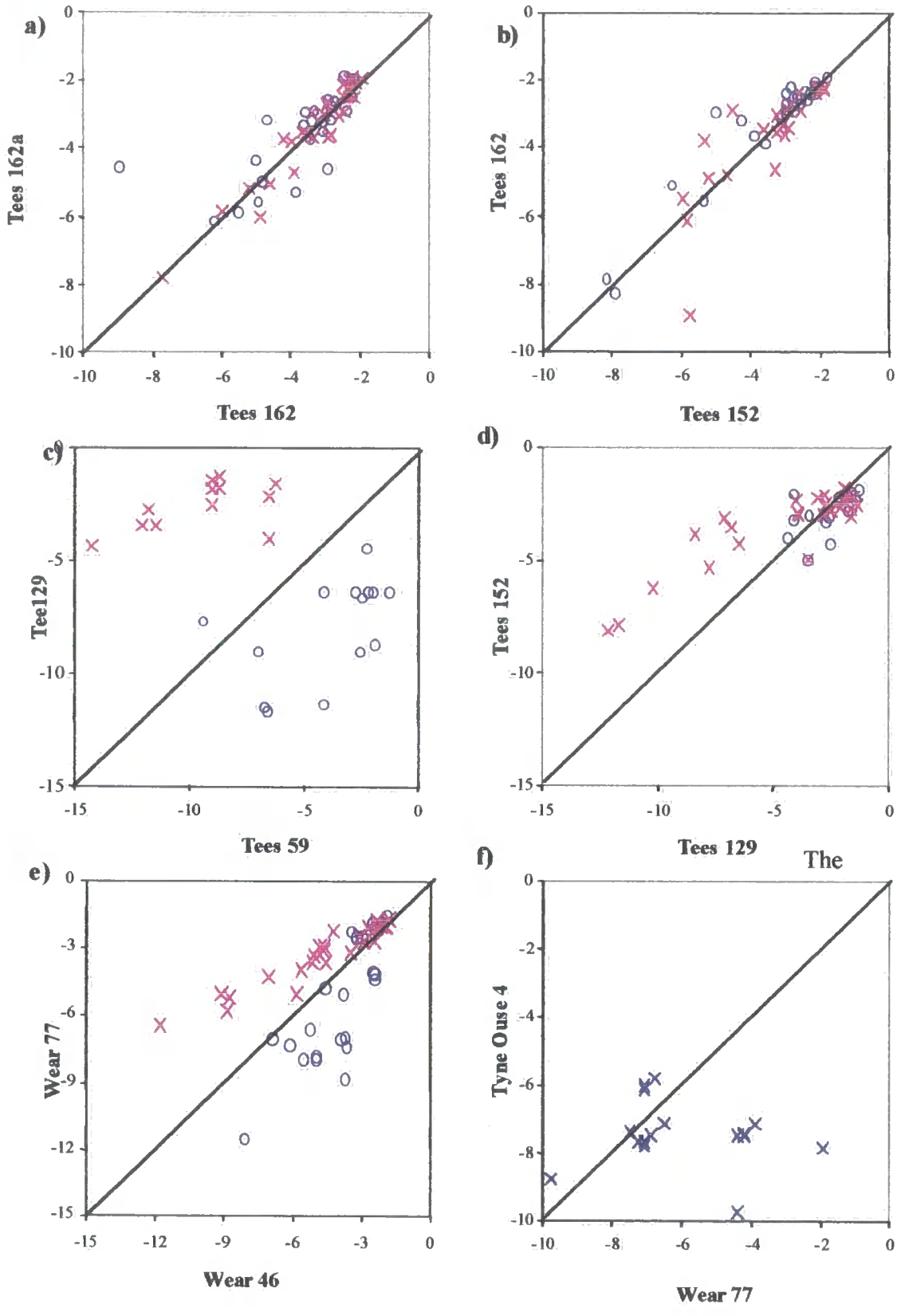
4.4.5 Chloroplast microsatellite variation

The mononucleotide repeat chloroplast microsatellite was analysed separately from the other microsatellites because it is haploid. The allele frequencies are shown in Table 4.1. All populations except Tyne Ouse 4 were monomorphic for an allele of 98 bp, but population Tyne Ouse 4 had three alleles of 97, 99 and 100 bp. Population Tyne Ouse 4 is found in a tributary of the Tyne and analysis of the other microsatellite loci have also found it to differ greatly from all other populations. It appears to have arisen from an independent introduction that originated from an original population that was from location that was separate from the other populations.

4.4.6 Assignment tests

Assignment tests revealed the similarities already seen between the most upstream three populations in the Tees (Figure 4.7a,b and d). However, there was overlap of just one genotype in the two most downstream populations in the Tees (Figure 4.7c), and this is in agreement with the tests of genetic variation carried out above.

Figure 4. 7 Genotype Assignment Tests. Maximum likelihood analysis of individual genotypes from each pair of populations originating from their own population versus the other population. Units are log likelihood values. Comparisons shown are: a) Tees 162 v Tees 162a; b) Tees 152 v Tees 162; c) Tees 59 v Tees 129; d) Tees 129 v Tees 152; e) Wear 46 v Wear 77; f) Maximum likelihood analysis of genotypes of population Tyne 94 being assigned to Tyne Ouse 4 versus Wear 77.



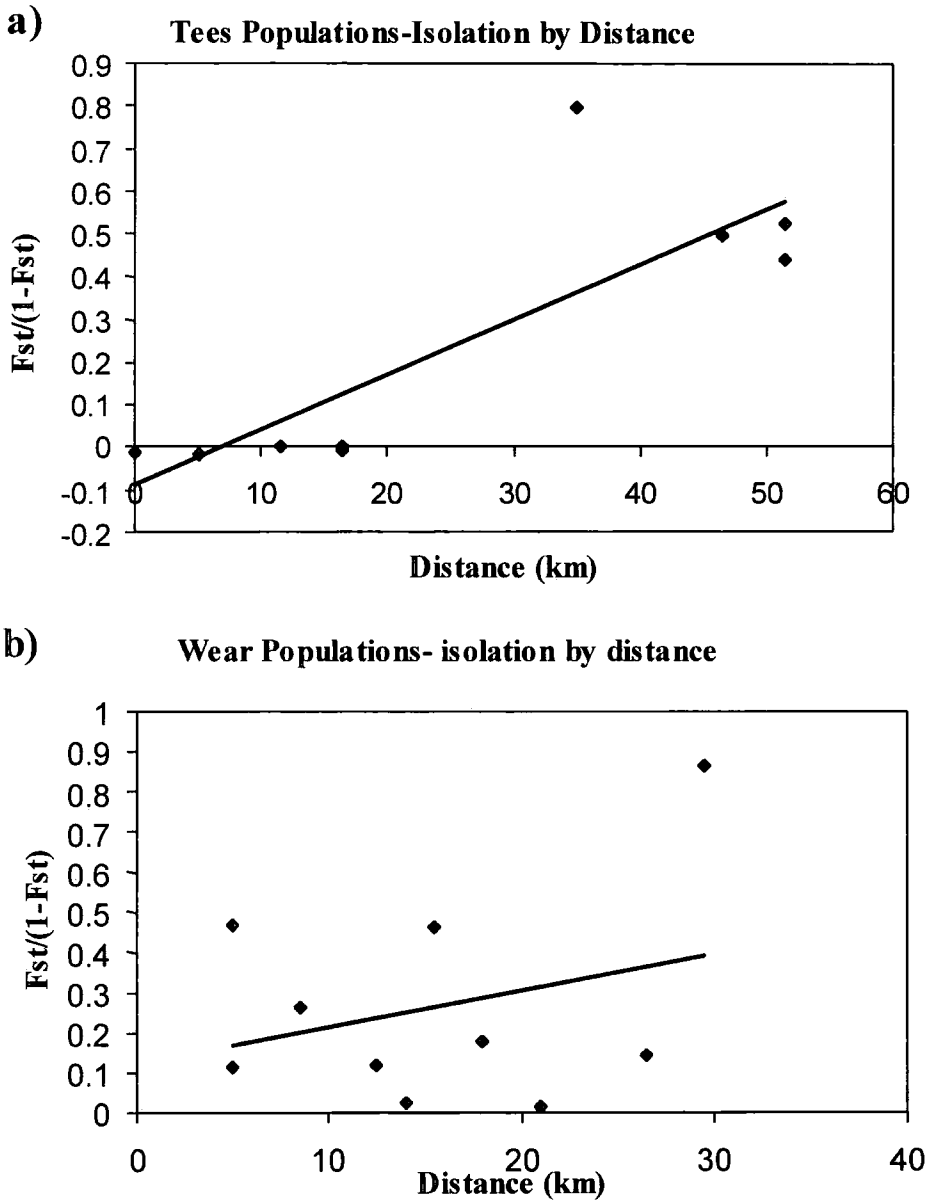
The populations compared in graphs 4.7d and 4.7e are similar distances apart from one another, but the populations in the Tees appear more similar than those in the Wear. This may occur because the populations in the Tees are found in a rural area but those in the Wear occur downstream of Durham city and so there is a greater chance that there has been more than one introduction into the Wear.

The likelihood of genotypes from population Tyne 94 originating from the other population in the Tyne (Tyne 94) or from a population in the Wear were plotted on Figure 4.9f. Genotypes are not obviously more similar to either population. This provides further evidence for the populations in the Tyne being very different from one another.

4.4.7 Isolation by distance

Comparison of the geographical and genetic distance between populations in the Tees (Figure 4.8a) shows there to be a significant positive correlation (P=0.002) with populations close together being very similar in terms of genetic variation, whilst populations further away were very different. However, if only populations more than 30 km apart are considered, the pattern of isolation by distance is not observed. Similar comparisons in the Wear did not show a positive correlation (Figure 4.8b).

Figure 4.8 Isolation by distance analysis of populations in the Tees and those in the Wear. Points are the results of pairwise comparisons between populations. Figure a): the equation for the best fit line is y= 0.0128x- 0.0865, R^2=0.709, *P*= 0.002. Figure b): the equation for the best fit line is y= 0.0002x+ 0.1892, R^2=0.085, *P*= 0.415.

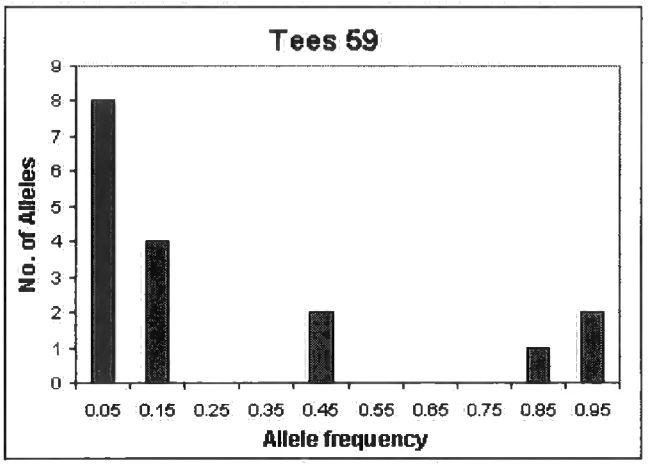


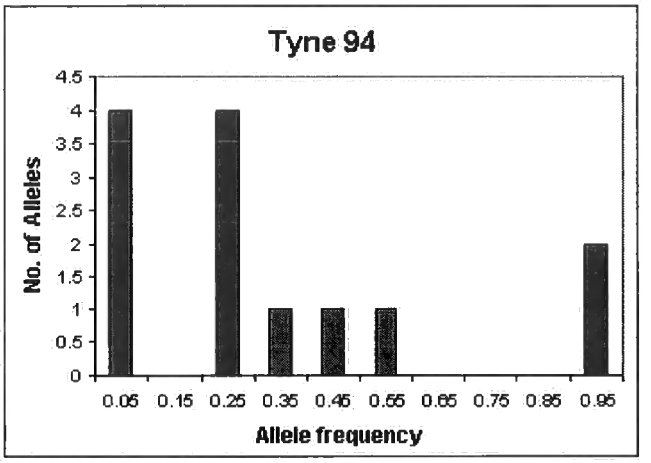
4.4.8 Detection of Bottlenecks

Luikart *et al.* (1997) presented a graphical method for detecting recent bottlenecks in populations (Section 3.11.3). The frequencies of alleles of multilocus genotypes in populations were placed into classes. The chart for population Tees 59 was typical of all but one of the other populations, with no suggestion that it had gone through a bottleneck as the lowest frequency class had the largest number of alleles. However, population Tyne 94 had the same number of alleles at an intermediate frequency as at

the lowest. This suggests that the population may have recently gone through a bottleneck. The population was relatively small and distant from other populations. The test was recommended to be carried out on at least five loci and here only four were considered and of these, one was monomorphic, so the test was in effect only carried out on three loci, which could bias results.

Figure 4.9 Number of alleles found in frequency classes of four microsatellite loci.





4.4.9 Genetic Distance Trees

Trees showing genetic distance were constructed using the PHYLIP phylogeny inference package (version 3.5) (section 3.12.).

The CONTML program was used to estimate genetic distance using a maximum likelihood approach. The trees produced using this program were redrawn using the DRAWTREE. In order to compare the genetic and geographic distances between populations, the DRAWTREE program was used to produce a tree which

would best fit onto a map of the study populations (Figure 4.10). The ends of the genetic distance tree refer to a population and the arrows join each population to its geographical location.

The four most upstream populations in the Tees were very similar genetically and all populations were very different from the Tees 59 population. The populations from the Wear were all present at the same end of the tree, with Wear 71 and Wear 35 on the end of one branch. Population Wear 18 and that from London were on the same branch, possibly because both have the greatest number of alleles. Both populations from the Tyne were on individual branches.

4.4.10 Construction of Trees based on Pairwise Genetic Distance Comparisons

The GENDIST program was used to compute genetic distance between populations and the results were used to produce trees using the FITCH, KITSCH and NEIGHBOUR programs (section 3.12.2).

The tree produced using FITCH was similar to that using CONTML, as in both cases populations in the Wear were grouped together (with the population from London located at the end of the same branch as population Wear 18), and population Tees 59 was separated from the other Tees populations, which were grouped together (Figure 4.11).

Figure 4.10 Distribution of populations of *H. mantegazzianum* and their genetic relatedness (section 4.4.9)

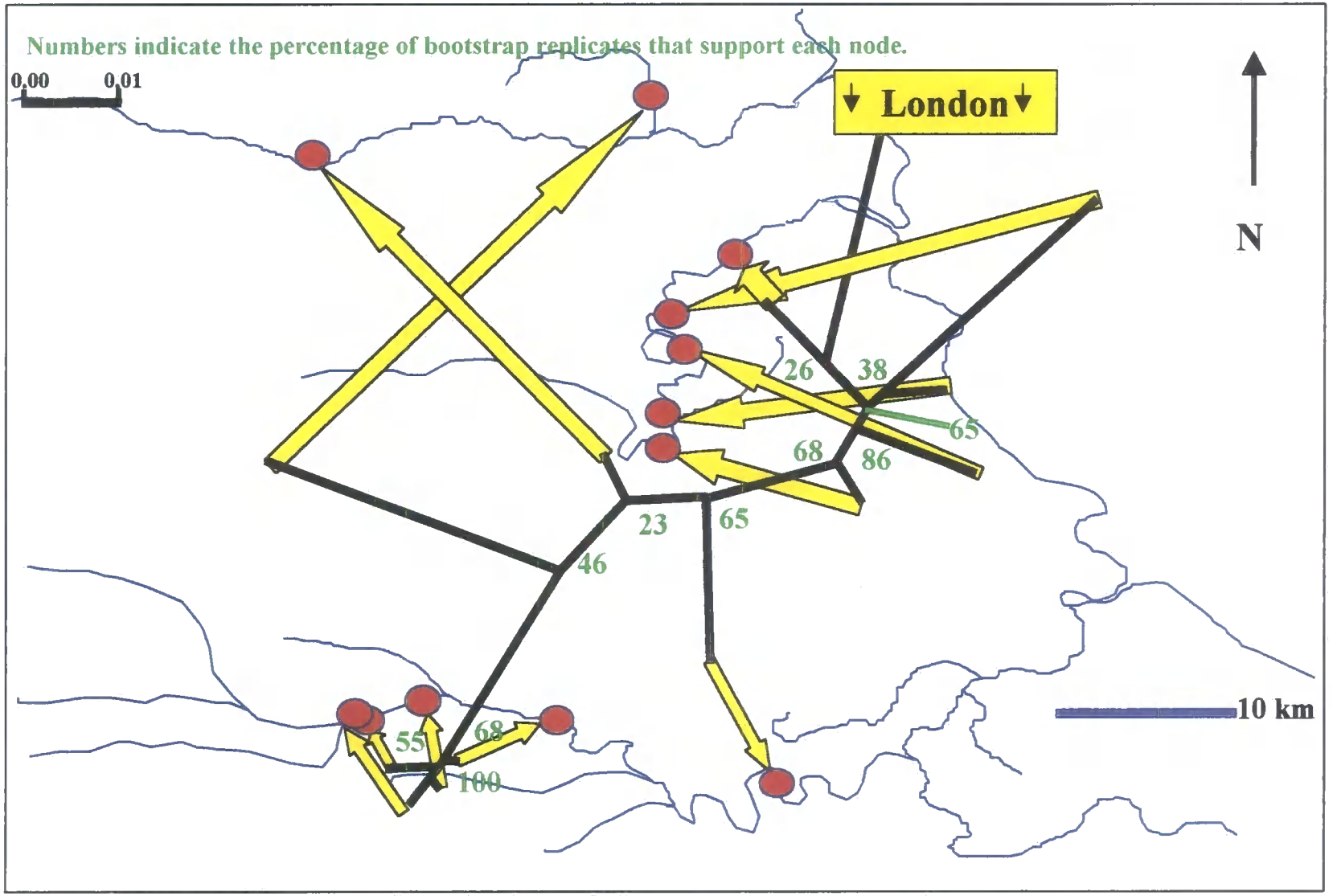
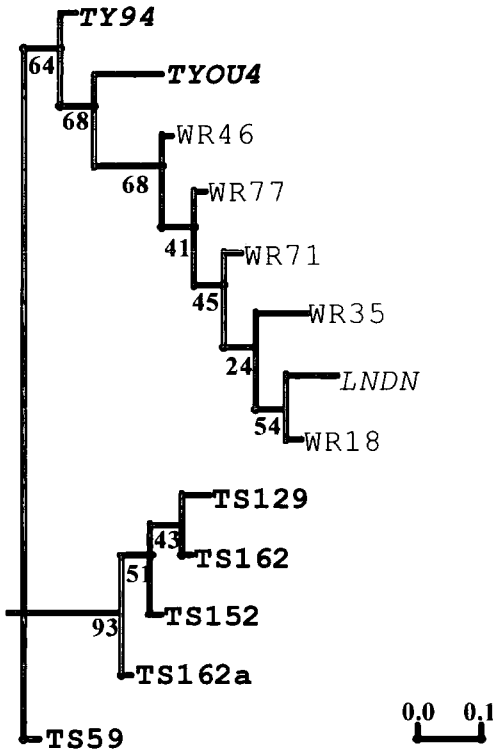


Figure 4.11 Tree produced using the FITCH program. Populations from the Tees are shown in bold, those from the Tyne in bold italics and the London population is in italics. Numbers indicate the percentage of bootstrap replicates that support each node.



KITSCH, which takes into account the differences in sizes of alleles, differed by placing population Tyne Ouse 4 separate from all other populations, and grouped population Tyne 94 with Tees 59 (Figure 4.12). The population from London was on a branch alongside the branch containing all populations from the Wear.

NEIGHBOR placed the population from London on the same branch as population Wear 18 (as with FITCH), but population Wear 35 was on its own branch, separate from the other Wear populations (Figure 4.13). All populations from the Tees shared a branch, but population Wear 59 was on a separate branch from the other Tees populations.

Figure 4.12 Tree produced using the KITSCH program. Populations from the Tees are shown in bold, those from the Tyne in bold italics and the London population is in italics. Numbers indicate the percentage of bootstrap replicates that support each node.

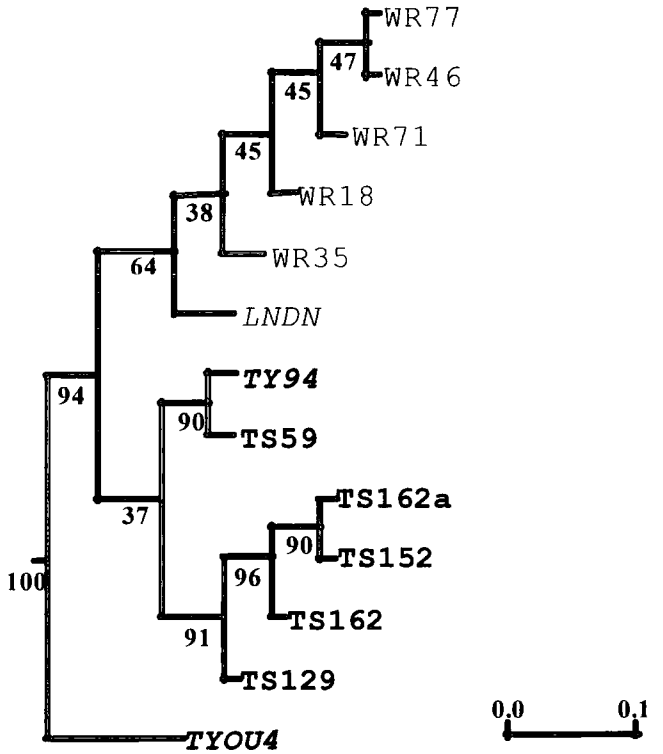
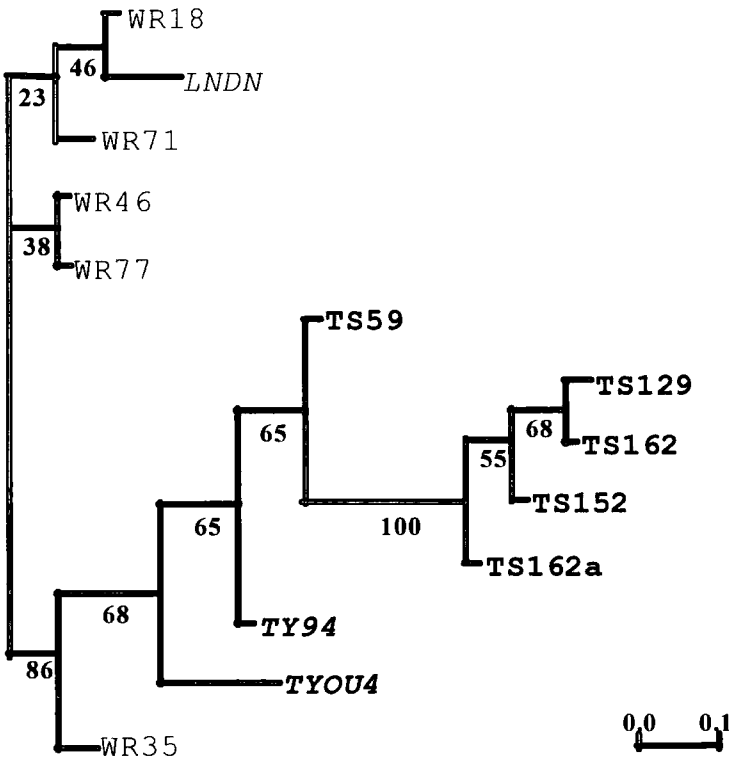


Figure 4.13 Tree produced using the NEIGHBOR program. Populations from the Tees are shown in bold, those from the Tyne in bold italics and the London population is in italics. Numbers indicate the percentage of bootstrap replicates that support each node.



4.5 Genetic Variation in the Tees

Populations Tees 162 and Tees 162a were sampled from the same site and F_{ST}, Rho_{ST} and maximum likelihood analysis found there to be no significant difference between them (section 4.4.2; Figure 4.7a). There was also no significant difference in F_{ST} values between population Tees 152 and populations Tees 162 and Tees 162a (Table 4.4). The similarity can also be seen in the assignment test (Figure 4.7b) comparing these populations and in genetic distance trees (Figures 4.10 and 4.11).

The Rho_{ST} values of the comparisons between population Tees 129 and those upstream of it were zero (Table 4.5). However, the values of F_{ST} were low (0.04-0.05) but significantly different from zero (Table 4.4). The allele frequency pie charts of populations in the upper Tees show similarities between all populations, although population Tees 129 was not as similar as the two most upstream populations (Figure 4.5; section 4.4.1). The second most common allele of microsatellite locus C52 in populations Tees 152, Tees 162 and Tees 162a was absent in population Tees 129. The assignment test comparing populations Tees 129 and Tees 152 support the inference of there being a much greater difference between these populations than populations Tees 152 and Tees 162 (Figure 4.7d). However, a number of individuals from population Tees 129 were more similar to population Tees 152 than to its own population, and vice versa.

Population Tees 59 was located 35 km downstream from population Tees 129. F_{ST} values comparing population Tees 59 with all other populations in the Tees ranged between 0.19 and 0.26 (Table 4.4). These values were much higher than such values between any other populations in the Tees. The difference between populations Tees 59 and Tees 129 was also apparent from the assignment test shown in Figure 4.7c where every individual was more similar to its own population. Population Tees 59 had five alleles found nowhere else in the Tees and many of the most common alleles in loci A34 and A43 for all upstream populations were absent from population Tees 59 (Table 4.1). There were very large Rho_{ST} values for comparisons of all other populations in the Tees with Tees 59, whereas all other Rho_{ST} values between populations in the Tees were zero (Table 4.5). The many differences between population Tees 59 and the others in the Tees suggest that this population has arisen from an independent introduction.

There was a pattern of isolation by distance observed in the Tees (Figure 4.8a). Apart from population Tees 59, all other populations in the Tees were grouped together in all four genetic distance trees (Figures 4.10 - 4.13).

4.6 Genetic Variation in the Wear

Populations Wear 77 and Wear 71 were only three km apart, but the F_{ST} value comparing these two populations was 0.12 and the difference was significant at the level of $P<0.05$ (Table 4.4). The Rho_{ST} value was even larger at 0.31 (Table 4.5). The second most common allele of locus A34 in population Wear 71 was absent from population Wear 77 and over the three other nuclear microsatellite loci, there were three alleles in population Wear 77 that were absent from population Wear 71 (Table 4.1).

Population Wear 46 was located 13.5 km downstream of population Wear 71 and the F_{ST} value comparing this population with population Wear 71 showed them to be significantly different, whilst populations Wear 46 and Wear 77 were not significantly different (Table 4.4). The Rho_{ST} value for the comparison of populations Wear 46 and Wear 77 was also very low at 0.01 (Table 4.5). Figure 4.6 shows the alleles of locus A34 for all populations. Population Wear 46 can be seen to appear to have similarities to both upstream populations.

Comparisons of F_{ST} values revealed population Wear 35 to be significantly different from all upstream populations (Table 4.4). At locus A34, the second most common allele present in population Wear 35 was not found in any of the upstream populations (Figure 4.6).

F_{ST} and Rho_{ST} values comparing population Wear 18 with those upstream show this population to be more similar to population Wear 46 than to population Wear 35 (Tables 4.4 and 4.5). At three microsatellite loci, population Wear 18 had alleles found nowhere else in the Wear, (Table 4.1) and the most common allele at locus A34 was found nowhere else. Three out of four genetic distance trees found population Wear 18 to be closest to the population from London (Figures 4.10-4.13).

4.7 Genetic Variation in the Tyne

The values of F_{ST} and Rho_{ST} comparisons between the two populations from the Tyne were relatively large and were similar to the values between these populations and all other populations (Tables 4.4 and 4.5). The genotype assignment test shown in Figure 4.7f shows population Tyne 94 to be as similar to a population in the Wear as to population Tyne Ouse 4. Considering all genomic microsatellite loci, population Tyne Ouse 4 had four alleles found nowhere else in the study area and the alleles found in the chloroplast microsatellite locus were not present in any other population (Table 4.1).

Chapter Five

Results II: *Impatiens glandulifera*

5.1 Populations, catchments and microsatellite loci

Thirteen populations of *I. glandulifera* were genotyped using the three polymorphic microsatellites identified from a genomic library and one chloroplast microsatellite (Chapter 3). There was no variation observed in the non-microsatellite cpDNA markers (section 3.3). The total number of individuals sampled was 390. The populations and their locations are shown in Figure 5.1. Seeds of *I. glandulifera* are available from garden centres in the study area and seeds from two packets produced by a grower in West Sussex (Vesutor Ltd.) and sold nationally, were genotyped. The names of the populations refer to the section on the rivers in which they occur as explained in section 4.1. Populations Tees 140 and Wear Rainton 52 were sampled in July 1997 and resampled as population Tees 140a and Wear Rainton 52a in May and June 1999 respectively.

The frequencies of alleles of the loci for each population are given in Table 5.1. The names of the loci are A2, A3, A21 and C2. C2 is a mononucleotide chloroplast microsatellite locus.

As was the case with *H. mantegazzianum*, a significant amount of genetic variation was found both within and between populations of *I. glandulifera*. The number of alleles found in the three genomic microsatellite loci varied from eight to sixteen. Eight alleles were found in one population for one locus. The number of alleles at each locus for each population is shown in Figure 5.2. The average number of alleles per locus per population was 4.7.

Figure 5.1 Distribution of populations of *I. glandulifera* in the study area

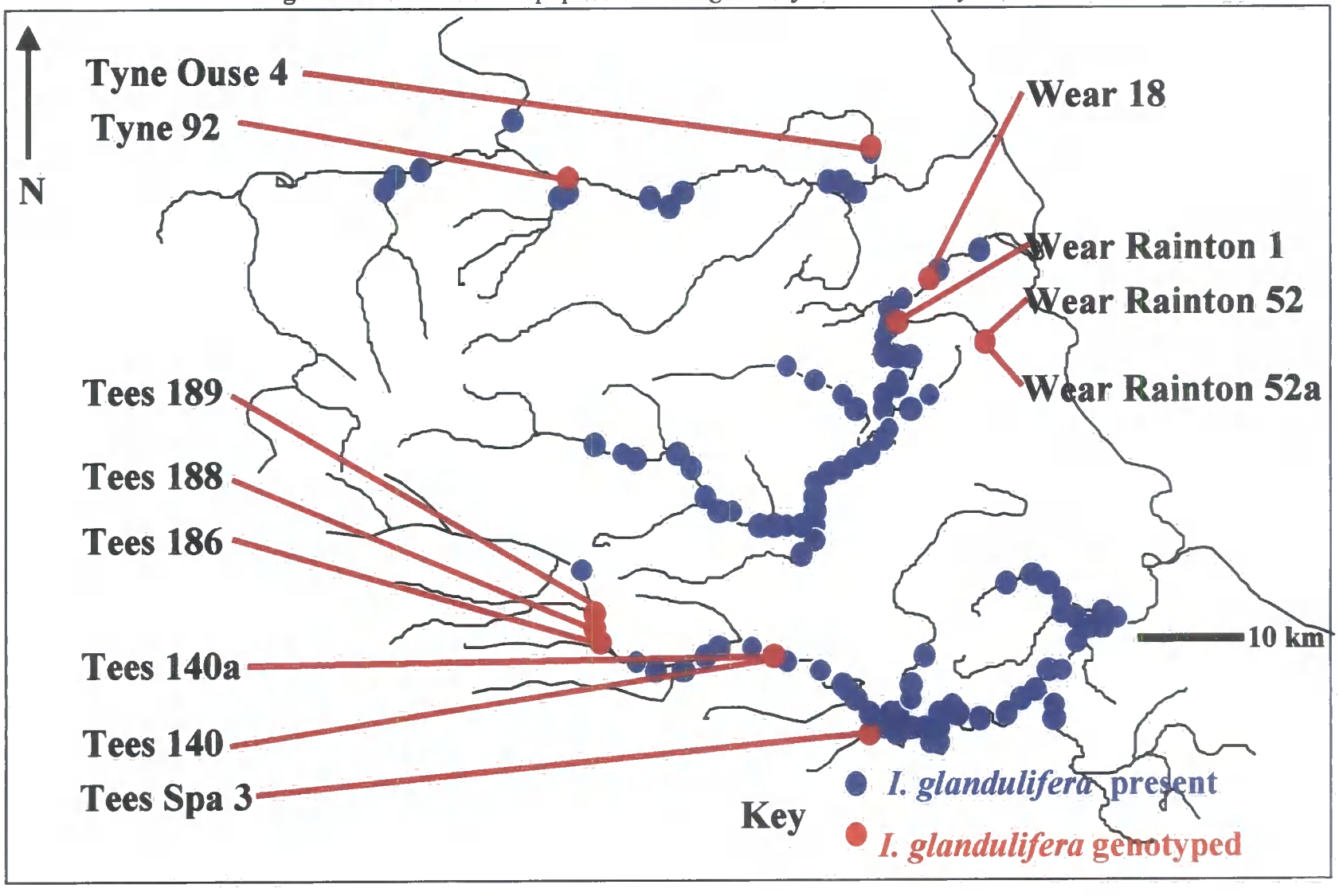
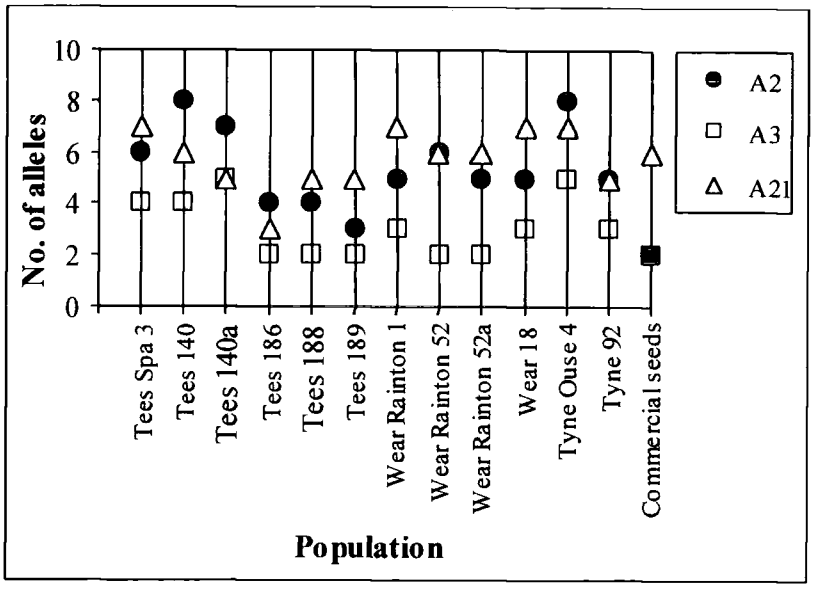


Table 5.1 Table of Allele Frequencies found in each population at each locus. Allele sizes are numbers of base pairs of the locus. Populations are grouped into catchments. See section 5.4 for population codes.

locus	size	TSp3	T140	T140a	T186	T188	T189	WR1	WR52	W52a	W18	TOU4	TY92	SEED	size
A2	**304**	2	2	2		1				1		8			**304**
	306	1		6	2	8	6	3		1	2	24			**306**
	308								1			11			**308**
	310								3			3	2		**310**
	312			4											**312**
	314								7	12		3			**314**
	316	16	14	2	1	19	22	24	46	44	29				**316**
	318	2	1						2			1	2		**318**
	320		1	1									1		**320**
	322		1					1							**322**
	324									2					**324**
	326											2			**326**
	328										1				**328**
	330	2	7	2	4			3			4		4	7	**330**
	332	37	33	43	53	32	32	29	1		24	8	51	53	**332**
	334		1												**334**

locus	size	TSp3	T140	T140a	T186	T188	T189	WR1	WR52	W52a	W18	TOU4	TY92	SEED	size
A3	**334**	6		2									3		**334**
	338		1												**338**
	340											1			**340**
	342							3				1			**342**
	344	2	2	1								1			**344**
	346			1							1				**346**
	348	22	44	41	54	32	35	36	55	39	41	43	28	43	**348**
	350	30	13	15	6	28	25	21	5	21	18	14	29	17	**350**
A21	**315**	22	18	25	38	21	14					8			**315**
	328										1				**328**
	332		1					1	18	8				2	**332**
	334	12		4		8	11	5	17	10	2	5	7	9	**334**
	336	15	17	19	18	18	15	24	13	20	31	10	2	6	**336**
	338	1	2					10	6	4	15	4			**338**
	340							3			1			4	**340**
	342												2		**342**
	354											1			**354**
	357	1													**357**
	359	2	13	10	4	5	6	13	4	13	5	25	15	21	**359**
	361	7	9	2		8	14	4	2	5	5	7	34	18	**361**
C2	**205**	7	12	14	21	15	14	7	30	30	8	13	22	0	**205**
	206	23	18	16	9	15	16	23	0	0	22	17	8	30	**206**

Figure 5.2 Comparison of number of alleles at each nuclear microsatellite locus in all populations.



5.2 Linkage Disequilibrium

The statistical tests carried out on microsatellite variation assume that all loci are independent of one another. Genepop 3.2 was used to test for linkage disequilibrium between the microsatellites (section 3.11.4). As none of the *P*-values were significant, the loci can be assumed to be independent of one another.

Table 5.2 Probability test of linkage disequilibrium for each locus pair across all populations (Fisher's method).

Locus pair	Chi^2	d.f.	*P-value*
A2 & A3	30.0	26	0.27
A2 & A21	32.7	26	0.17
A3 & A21	35.9	26	0.09

5.3 Catchment Analysis

In order to compare genetic variation between catchments, all populations in each catchment were pooled.

5.3.1 F_{ST} values of catchment comparisons

F_{ST} values were computed using the Arlequin program, the results of which can be seen in the tables below.

Table 5.3 Matrix of F_{ST} values for catchment comparisons. Values were all significant differences at the level of $P<0.05$.

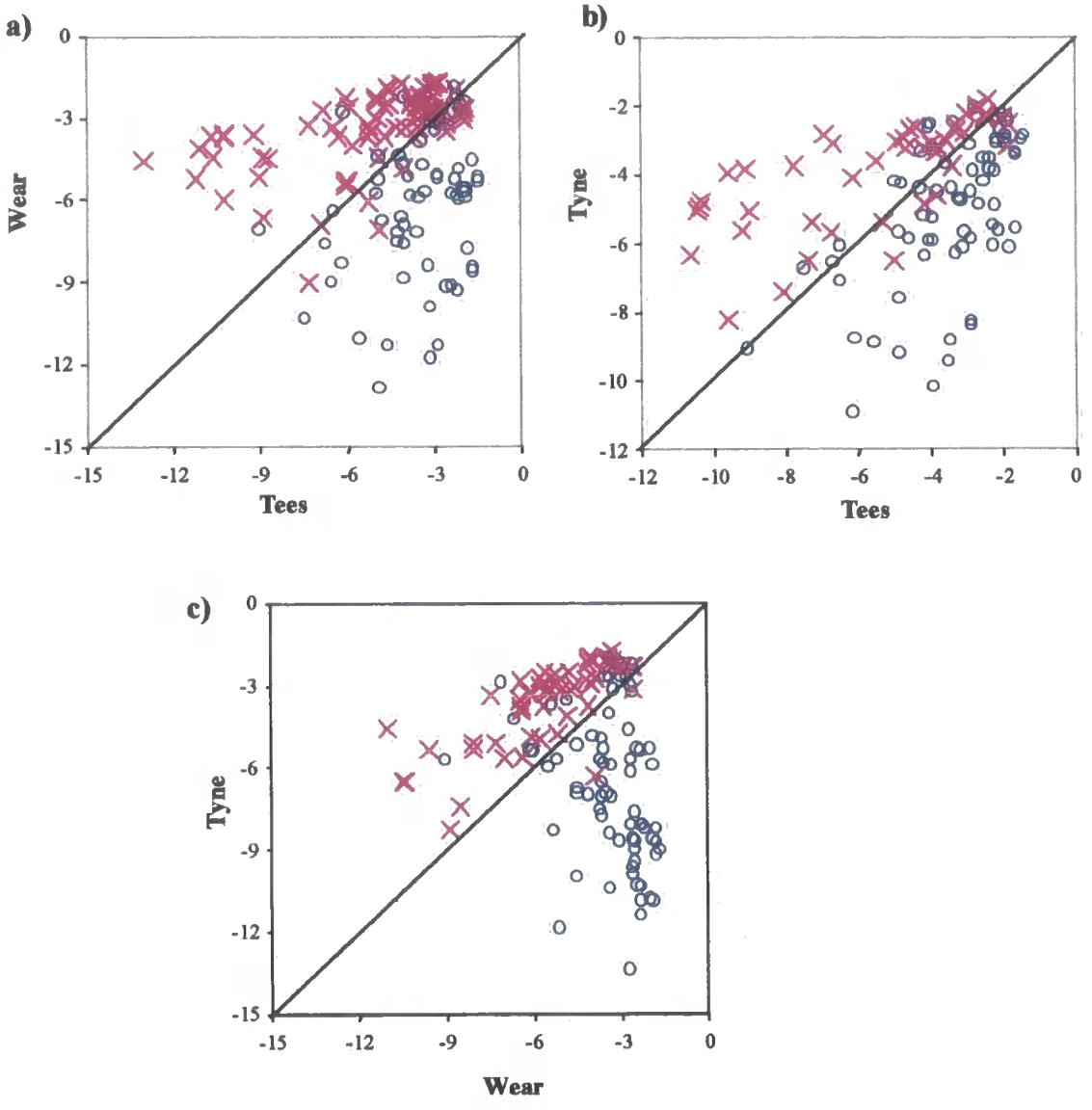
	Tees	Wear
Wear	0.13	
Tyne	0.08	0.16

There were significant differences between all catchments, which was expected considering that the species is thought to be unable to disperse between catchments without the aid of humans.

5.3.2 Assignment tests

Figure 5.3 shows pairwise plots of maximum likelihood. Any genotypes that cross the diagonal line from other genotypes from the same catchment are more similar to the other catchment than to their own. The separation between individuals from the Tyne and Wear catchments can be clearly seen, but there is quite a lot of overlap in the comparison between the Tees and the Wear and a large amount of overlap between the Tees and the Tyne.

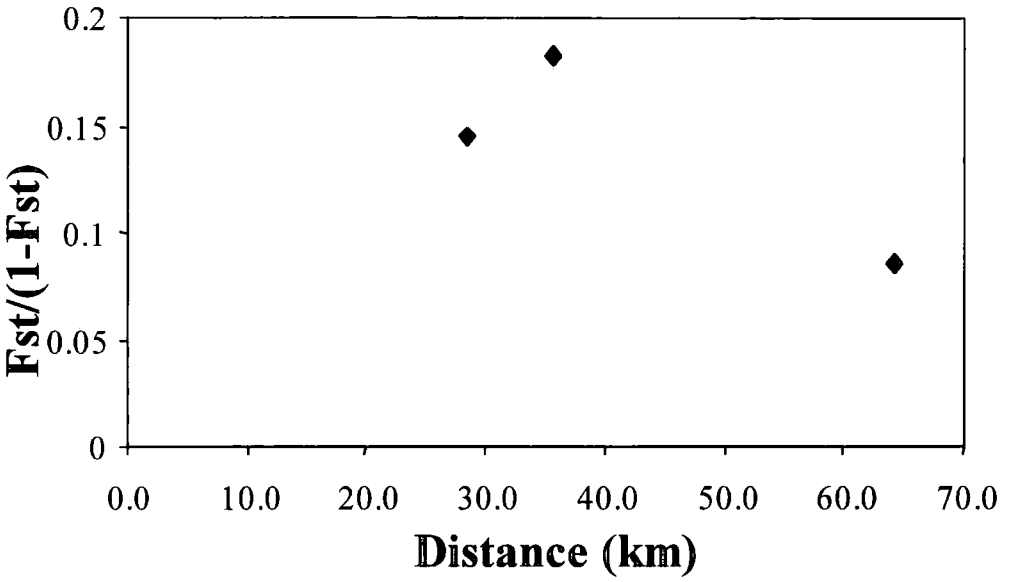
Figure 5. 3 Genotype Assignment Tests. Maximum likelihood analysis of individual genotypes from each pair of catchments originating from their own catchment versus the other catchment. Comparisons shown are a) Tees v Wear, b) Tees v Tyne, c) Wear v Tyne. Units are log likelihood values of individuas belonging to each population.



5.3.3 Isolation by distance

The distance between catchments was calculated by calculating the midpoint of the distribution of the species in the Tees, Tyne and Wear river catchments. The distance between these was measured.

Figure 5.4 Isolation by distance analysis of catchments. Points are the results of pairwise comparisons between *I. glandulifera* from the Tees, Tyne and Wear river catchments. Distance between catchments is the distance between the midpoint along each main river through which plants were sampled.



Plots of F_{ST} /(1- F_{ST}) against distance and ln distance (Figure 5.4) show no relationship, which was not unexpected, given the paucity in number of comparisons. However, the lowest value of F_{ST} /(1- F_{ST}) was found for the comparison of the Tees and Tyne, which were the most geographically distant catchments.

5.4 Population analysis

In the following figures and tables, the populations were coded as follows:

Tees Spa 3-	TSp3	Tees 189-	T189	Tyne Ouse 4-	TOU4
Tees 140-	T140	Wear Rainton 1-	WR1	Tyne 92-	TY92
Tees 140a-	T140a	Wear Rainton 52-	WR52	Commercial Seeds-	SEED
Tees 186-	T186	Wear Rainton 52a-	W52a		
Tees 188-	T188	Wear 18-	W18		

5.4.1 Pie Charts of allele frequencies

Figure 5.5 shows pie charts of the allele frequencies of four loci for populations Tees 140, 186, 188 and 189. Populations Tees 188 and 189 were very similar, as would be expected since they were only 0.5 km apart. Population Tees 186 was just one km downstream of population Tees 188 but was quite different in terms of allele frequencies. Population Tees 140 was a further 24 km downstream but had a number of alleles in all loci that were not observed in the upstream populations.

Figure 5.6 shows pie charts for locus A21 for ten populations. The dark blue allele was the most common allele found in the Tees and was present in all populations in the Tees, but the only population outside the Tees it was found in was Tyne Ouse 4. Population Wear Rainton 1, which was at the mouth of a tributary of the Wear was more similar to population Wear 18, on the main river of the Wear, than to population Wear Rainton 52 which was 25.5 km further upstream along the same tributary. The populations in the Tyne were quite different in the allele frequencies and the alleles present. Population Tyne Ouse 4 had four alleles not found in Tyne 92.

Figure 5.5 Distribution of alleles (coloured segments) of three microsatellite loci of *I. glandulifera* in four populations in the Upper Tees. Population codes are given in section5.4.

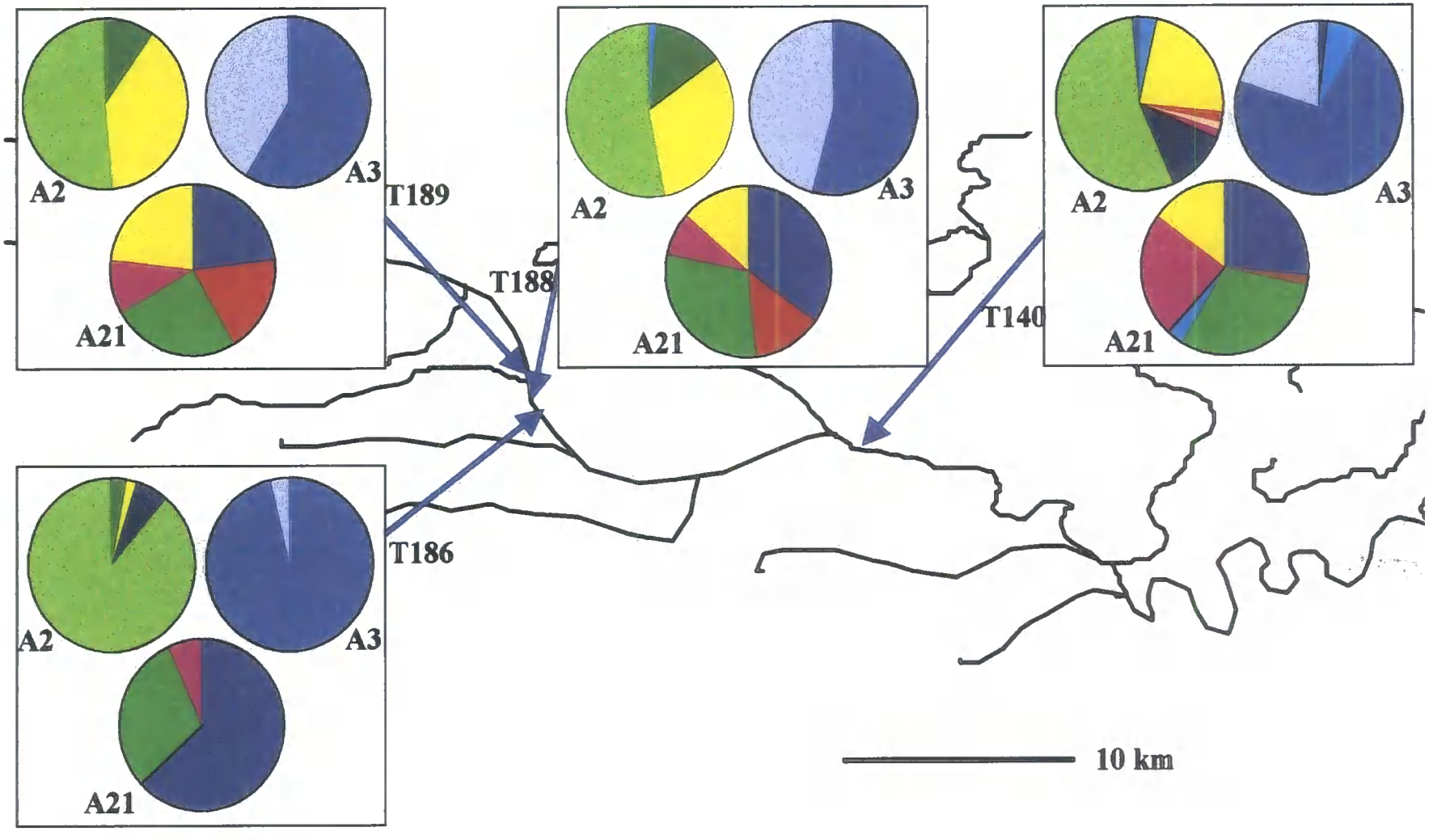
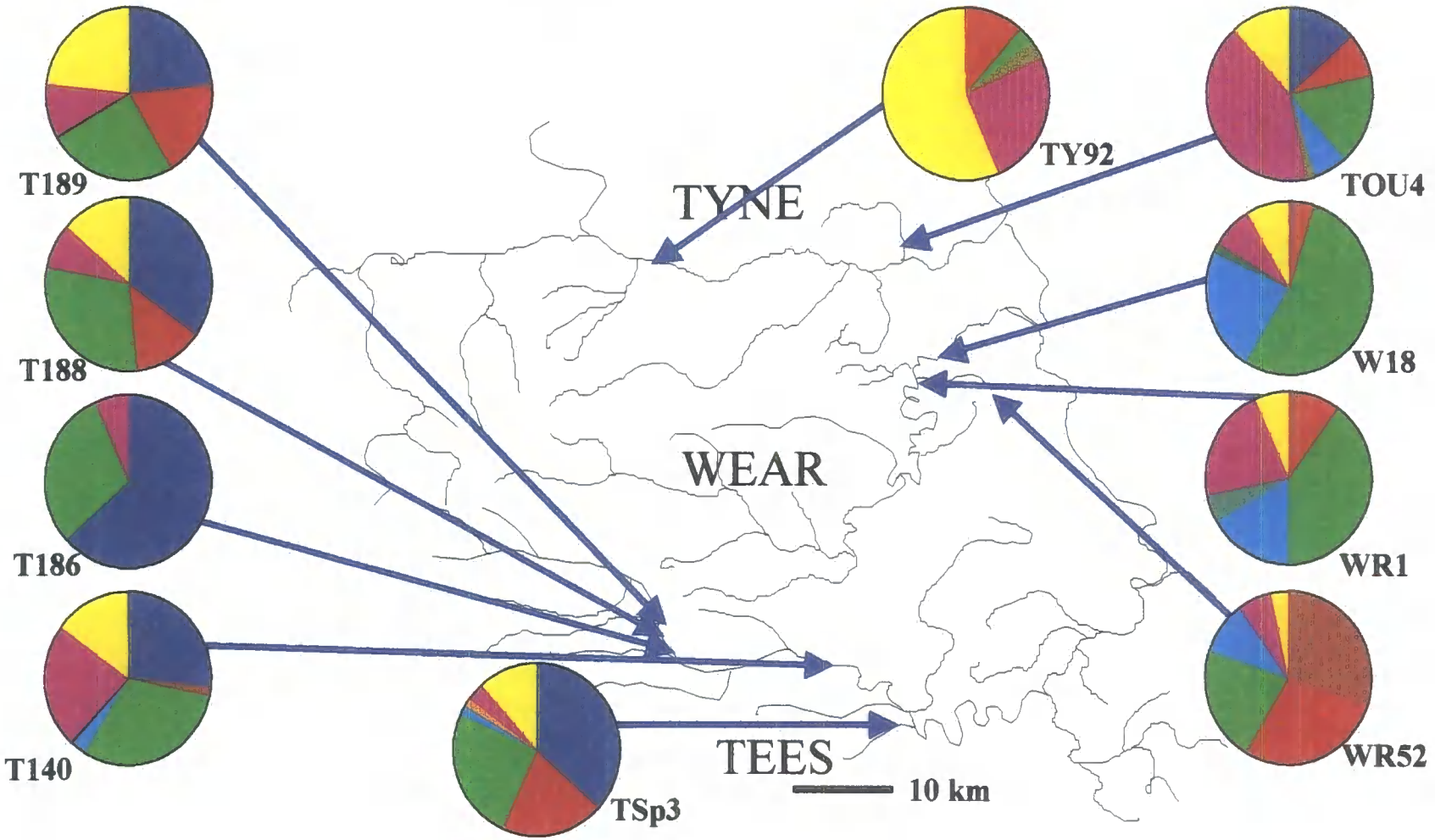


Figure 5.6 Distribution of different alleles (coloured segments) of locus A21 in ten populations of *I. glandulifera* in the Northeast of England. The population codes are given in section 5.4.



5.4.2 F_{ST} values of population comparisons

Table 5.4 Matrix of F_{ST} values across all loci. Population codes are given in section 5.4. Values in italics are non-significant differences at the level of $P<0.05$. One thousand permutations were carried out.

	TSp3	T140	T140a	T186	T188	T189	WR1	WR52	W52a	W18	TOU4	TY92
T140	**0.06**											
T140a	**0.06**	**0.02**										
T186	**0.19**	**0.10**	**0.05**									
T188	***0.00***	**0.03**	**0.04**	**0.17**								
T189	***0.01***	**0.02**	**0.06**	**0.19**	***-0.01***							
WR1	0.07	0.04	0.09	0.24	0.04	*0.02*						
WR52	0.28	0.22	0.32	0.46	0.24	0.20	**0.17**					
W52a	0.20	0.17	0.27	0.42	0.15	0.13	**0.09**	**0.06**				
W18	0.11	0.06	0.14	0.28	0.07	0.06	***0.00***	**0.15**	**0.09**			
TOU4	0.18	0.11	0.14	0.29	0.13	0.13	0.12	0.25	0.19	0.16		
TY92	0.13	0.15	0.17	0.32	0.15	0.12	0.17	0.43	0.34	0.24	**0.24**	
SEED	0.15	0.09	0.10	0.23	0.14	0.11	0.13	0.38	0.32	0.20	0.19	0.05

The tables above show that there is a large amount of variation between populations. Eighty-two percent of F_{ST} values for comparisons of populations from different catchments were higher than 0.1, whereas only 32 % of F_{ST} values of comparisons of populations from the same catchment were higher than 0.1. Therefore, there was considerably more variation between than within catchments. Four out of the five population comparisons that were not significantly different were for populations in the same catchment.

5.4.3 Rho $_{ST}$ and $(\delta\mu)^2$ population comparisons

Rho $_{ST}$ values are given in the matrices below for individual loci and for all loci pooled together using the software program RST Calc. As in section 5.4.2, within population comparisons are shown in bold.

Table 5.5 Matrix of *Rho* $_{ST}$ comparisons. Population codes are given in section 5.4.

	TSp3	T140	T140a	T186	T188	T189	WR1	WR52	W52a	W18	TOU4	TY92
T140	**0.01**											
T140a	**-0.01**	**0.01**										
T186	**0.10**	**0.14**	**0.07**									
T188	**0.03**	**0.07**	**0.04**	**0.24**								
T189	**0.05**	**0.04**	**0.05**	**0.27**	**0.00**							
WR1	0.08	0.01	0.08	0.27	0.07	0.01						
WR52	0.20	0.20	0.18	0.41	0.21	0.21	**0.20**					
W52a	0.27	0.23	0.25	0.51	0.16	0.13	**0.15**	**0.09**				
W18	0.07	0.03	0.07	0.31	0.04	0.00	**0.00**	**0.19**	**0.14**			
TOU4	0.28	0.22	0.27	0.42	0.25	0.21	0.17	0.12	0.08	0.19		
TY92	0.27	0.18	0.23	0.25	0.29	0.24	0.24	0.52	0.44	0.31	**0.32**	
SEED	0.28	0.19	0.24	0.27	0.30	0.23	0.21	0.53	0.40	0.28	0.30	0.04

Populations Wear Rainton 52 and 52a, which were taken from the same location in different years, were very different from all other populations. The two populations themselves were signficantly different from one another, which was likely to have been due to the relatively small number of individuals at the location and because the species is an annual. This location was 26 km up a tributary. Population Wear Rainton 1, which was located at the mouth of the tributary, was more similar to populations in the Tees than to the populations in the same tributary.

Table 5.6 Matrix of $(\delta\mu)^2$ values. Population codes are given in section 5.4.

	TSp3	T140	T140a	T186	T188	T189	WR1	WR52	W52a	W18	TOU4	TY92
T140	**15**											
T140a	**0**	**15**										
T186	**27**	**73**	**25**									
T188	**4**	**12**	**5**	**49**								
T189	**18**	**3**	**19**	**87**	**9**							
WR1	40	8	41	127	28	5						
WR52	41	41	45	123	21	23	**37**					
W52a	70	45	74	182	42	25	**24**	**11**				
W18	24	4	25	100	13	0	**3**	**23**	**21**			
TOU4	140	99	146	286	99	70	60	39	12	62		
TY92	203	108	203	336	187	116	74	204	141	105	**160**	
SEED	140	65	138	245	132	76	46	163	118	69	153	8

$(\delta\mu)^2$ was calculated using RST Calc. as in section 3.11.1. The average $(\delta\mu)^2$ value for within catchment comparisons was 29, whereas the average $(\delta\mu)^2$ value for between catchment comparisons was 87.

5.4.4 Test of Hardy-Weinberg Equilibrium

The results of tests of deviation from the Hardy-Weinberg equilibrium are shown in Table 5.7. Significant heterozygote deficiencies were found in all loci.

Table 5.7 Test of conformity of heterozygosity levels to the Hardy-Weinberg equilibrium. Values in bold are significant deviations at α=0.05 after Bonferroni correction for Type I errors. #Indiv= Number of individuals; Obs. Heter= observed heterozygosity; Exp. Heter= expected heterozygosity.

Population	Locus	#Genot	Est. Fis	Obs.Heter.	Exp.Heter.	*P* value	s.d.
Tees Spa 3	A2	30	0.16	0.47	0.55	0.1723	0.0012
	A3	30	0.134	0.53	0.61	0.1486	0.0011
	A21	30	0.036	0.73	0.77	0.6724	0.0013
Tees 140	A2	30	0.272	0.47	0.67	**0.0001**	0.0000
	A3	30	0.211	0.33	0.45	0.3022	0.0012
	A21	30	0.094	0.70	0.77	0.1869	0.0010
Tees 140a	A2	30	0.444	0.27	0.52	**0.0000**	0.0000
	A3	30	0.163	0.40	0.48	0.0506	0.0005
	A21	30	0.007	0.70	0.72	0.0061	0.0003
Tees 186	A2	30	0.237	0.17	0.46	0.2261	0.0014
	A3	30	0.275	0.13	0.43	0.2425	0.0013
	A21	30	-0.04	0.53	0.51	0.1039	0.0010
Tees 188	A2	30	0.29	0.43	0.64	0.0316	0.0005
	A3	30	0.079	0.47	0.51	0.7212	0.0014
	A21	30	-0.012	0.77	0.76	0.0213	0.0005
Tees 189	A2	30	0.257	0.43	0.60	0.0701	0.0008
	A3	30	0.397	0.30	0.49	0.0575	0.0007
	A21	30	0.167	0.67	0.81	0.1368	0.0010
Wear Rainton 1	A2	30	0.568	0.27	0.63	**0.0003**	0.0001
	A3	30	0.303	0.37	0.55	0.1827	0.0010
	A21	30	0.041	0.73	0.78	0.2332	0.0012
Wear Rainton 52	A2	30	0.836	0.07	0.43	**0.0000**	0.0000
	A3	30	-0.074	0.17	0.19	1.0000	0.0000
	A21	30	-0.113	0.87	0.79	0.4000	0.0014
Wear Rainton 52a	A2	30	0.769	0.10	0.45	**0.0000**	0.0000
	A3	30	0.065	0.43	0.48	1.0000	0.0000
	A21	30	-0.002	0.80	0.81	0.2979	0.0012
Wear 18	A2	30	0.404	0.37	0.65	0.0086	0.0002
	A3	30	0.188	0.37	0.48	0.5649	0.0019
	A21	30	0.202	0.53	0.67	0.0776	0.0008
Tyne Ouse 4	A2	30	0.318	0.53	0.78	**0.0000**	0.0000
	A3	30	-0.143	0.50	0.44	0.8631	0.0008
	A21	30	0.046	0.73	0.78	0.4421	0.0013
Tyne 92	A2	30	0.398	0.17	0.36	0.0058	0.0002
	A3	30	0.524	0.27	0.60	**0.0007**	0.0001
	A21	30	0.239	0.47	0.66	0.0857	0.0007
Commercial seeds	A2	30	0.208	0.17	0.27	0.3269	0.0015
	A3	30	0.601	0.17	0.42	0.0025	0.0002
	A21	30	-0.274	0.97	0.76	**0.0000**	0.0000

5.4.5 Chloroplast microsatellite variation

The mononucleotide repeat chloroplast microsatellite was analysed separately from the other microsatellites because it is haploid.

Table 5.8 Matrix of F_{ST} values for the chloroplast locus C2. Non-significant differences at $P<0.05$ are shown in italics.

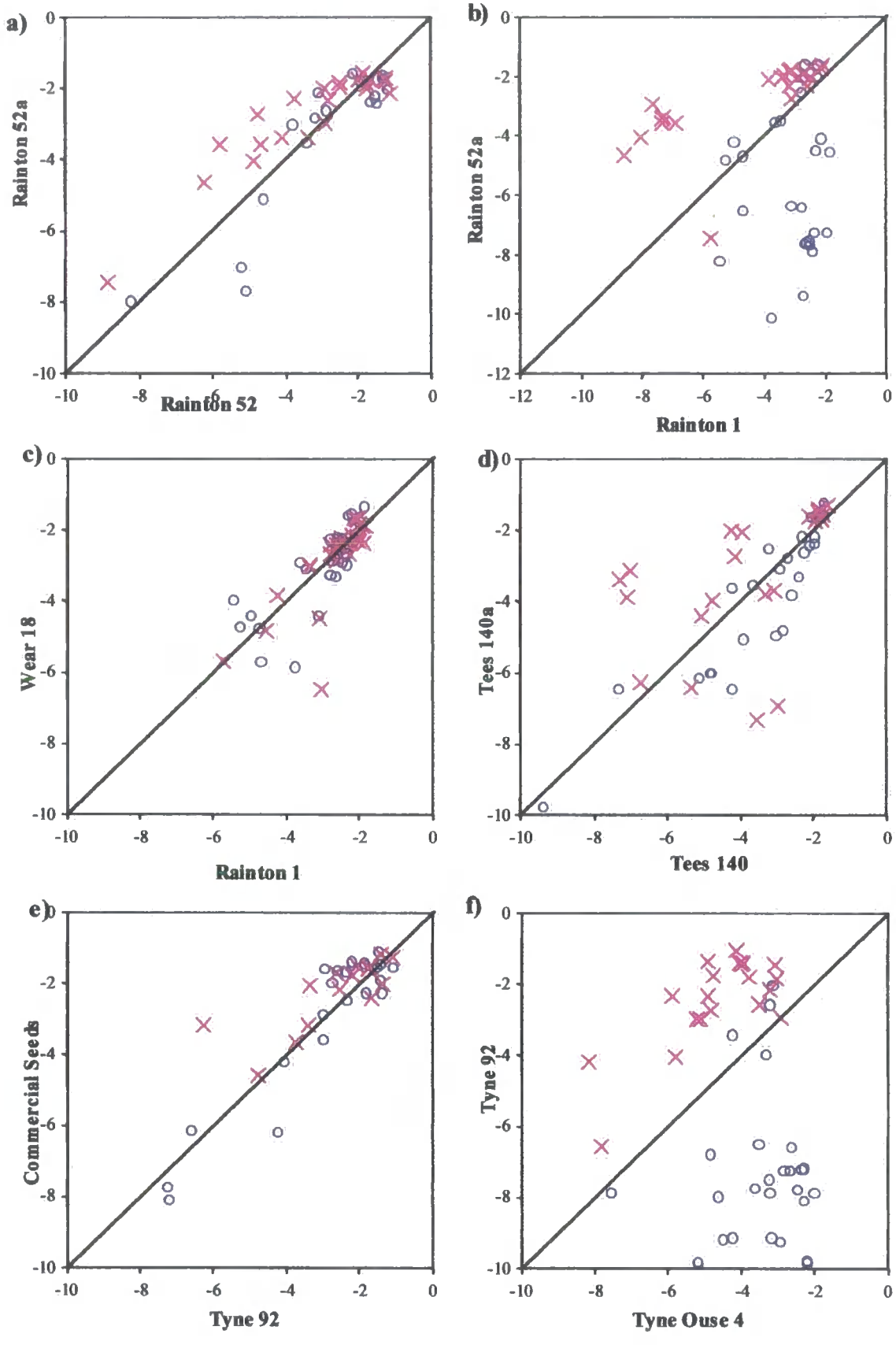
	TSp3	T140	T140a	T186	T188	T189	WR1	WR52	W52a	W18	TOU4	TY92
T140	*0.05*											
T140a	**0.10**	*-0.01*										
T186	**0.35**	**0.15**	*0.09*									
T188	*0.13*	*0.00*	*-0.01*	*0.06*								
T189	**0.10**	*-0.01*	*-0.02*	*0.09*	*-0.01*							
WR1	*-0.02*	*0.05*	0.10	0.35	*0.13*	0.10						
WR52	0.76	0.59	0.53	0.29	0.49	0.53	**0.76**					
W52a	0.76	0.59	0.53	0.29	0.49	0.53	**0.76**	*0.00*				
W18	*-0.01*	*0.02*	*0.07*	0.30	*0.09*	*0.07*	*-0.01*	**0.73**	**0.73**			
TOU4	*0.07*	*-0.01*	*-0.01*	*0.12*	*-0.01*	*-0.01*	*0.07*	0.56	0.56	*0.04*		
TY92	0.39	0.19	*0.12*	*-0.01*	*0.09*	*0.12*	0.39	0.25	0.25	0.35	**0.16**	
SEED	0.22	0.39	0.46	0.69	0.49	0.46	0.22	1.00	1.00	0.25	0.42	0.73

Only two alleles were identified. Populations Wear Rainton 52, Wear Rainton 52a and the Commercial Seeds were monomorphic, but Commercial Seeds was monomorphic for a different allele than the other two populations. Population Wear Rainton 1 was significantly different from populations Wear Rainton 52 and 52a, but was not significantly different from population Wear 18.

5.4.6 Assignment tests

Figure 5.7 shows pairwise plots of log-likelihood. Graph a) shows a comparison of the same location sampled two years apart. Graphs b) and c) compare Wear Rainton 1 with a population 26 km upstream on the same tributary and with a population along the main river (Wear 18). Wear Rainton 1 was much more similar to Wear 18 than to the population in the same tributary, as was seen from comparisons with F_{ST} values. Graphs e) and f) show that the two populations from the Tyne were very different. Population Tyne 92 was much more similar to the commercial seeds than the other population from the Tyne.

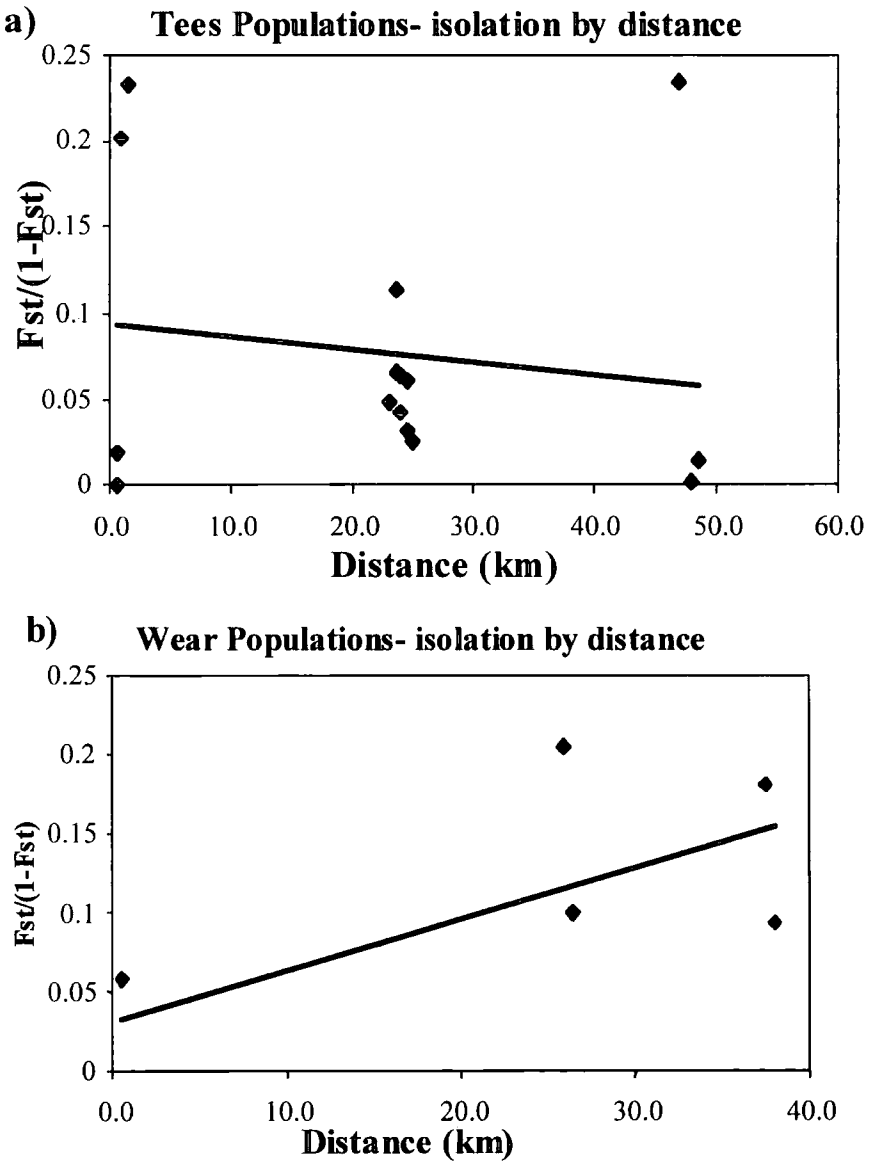
Figure 5.7 Genotype Assignment Tests. Maximum likelihood analysis of individual genotypes from each pair of populations originating from their own population versus the other population. Units are log likelihood values of individuas belonging to each population.



5.4.7 Isolation by distance

Figure 5.8 shows graphs comparing geographical and genetic distance between populations in the Tees and Wear. Graph b) shows a positive relationship between geographical and genetic distance (P=0.006). However, graph a) shows that no such relationship was found in the Tees as there was a negative relationship. This is partly due to population Tees 186 being geographically very close to populations Tees 188 and Tees 189, but quite different in terms of F_{ST}.

Figure 5.8 Isolation by distance analysis of populations in the Tees and those in the Wear. Points are the results of pairwise comparisons between populations. Figure a): equation of the best fit line was y= -0.0008x + 0.0942, R^2= 0.0245, P= 0.045. Figure b): equation of the best fit line was y= 0.0033x + 0.0313, R^2= 0.415, P= 0.006.



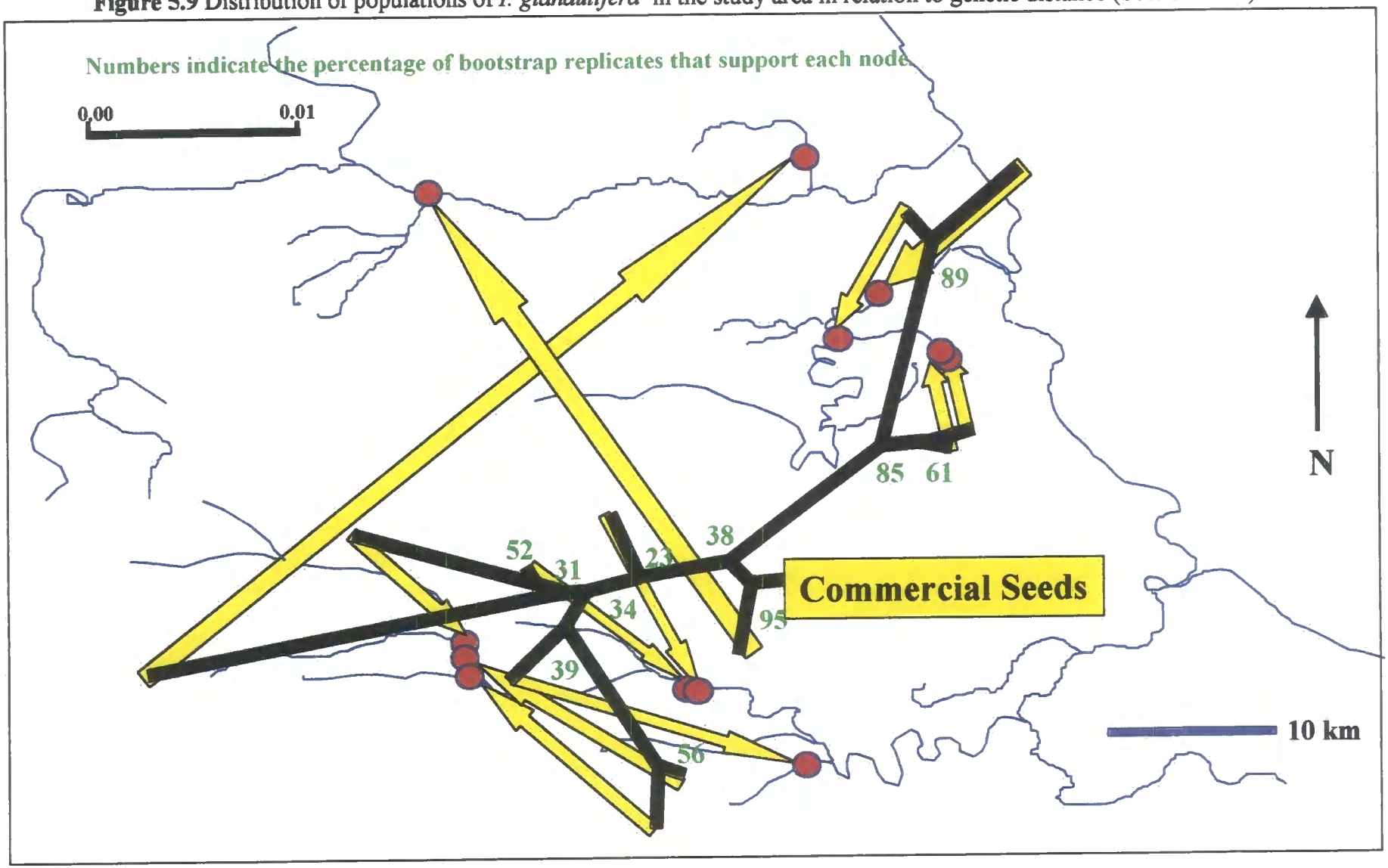
5.4.8 Genetic Distance Trees

Trees showing genetic distance were constructed using the PHYLIP phylogeny inference package (version 3.5) (section 3.12).

In order to compare the genetic and geographic distances between populations, the CONTML and DRAWTREE programs were used (section 4.4.9) to produce a tree which would best fit onto a map of the study populations (Figure 5.9). The ends of the genetic distance tree refer to a population and the arrows join each population to its geographical location.

The map overall shows less of a correlation between genetic and geographic distance compared which the equivalent map for *H. mantegazzianum* (Figure 4.10). Population 186 was on a different branch from populations 188 and 189, despite their very close proximity. Tees 140 and Tees 140a were sampled from the same location two years apart but were on different (but neighbouring) branches of the tree. The two populations sampled from the same location in the Wear, however, were close together on the same branch, as were the two most downstream populations. As was the case with populations of *H. mantegazzianum*, the two populations in the Tyne were on different branches. Population Tyne 92 was on the same branch as the population consisting of seeds commercially available.

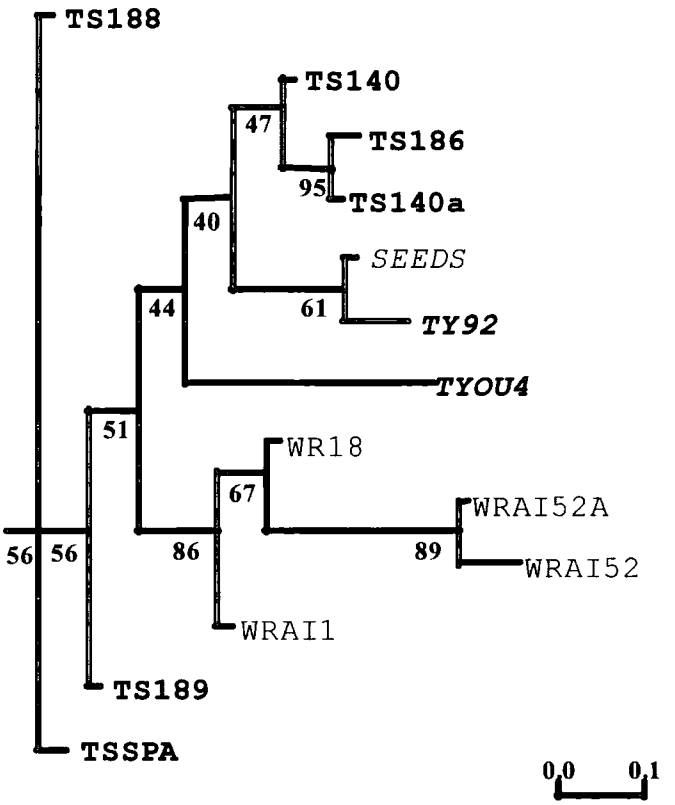
Figure 5.9 Distribution of populations of *I. glandulifera* in the study area in relation to genetic distance (Section 5.4.8)



5.4.9 Construction of Trees based on Pairwise Genetic Distance Comparisons

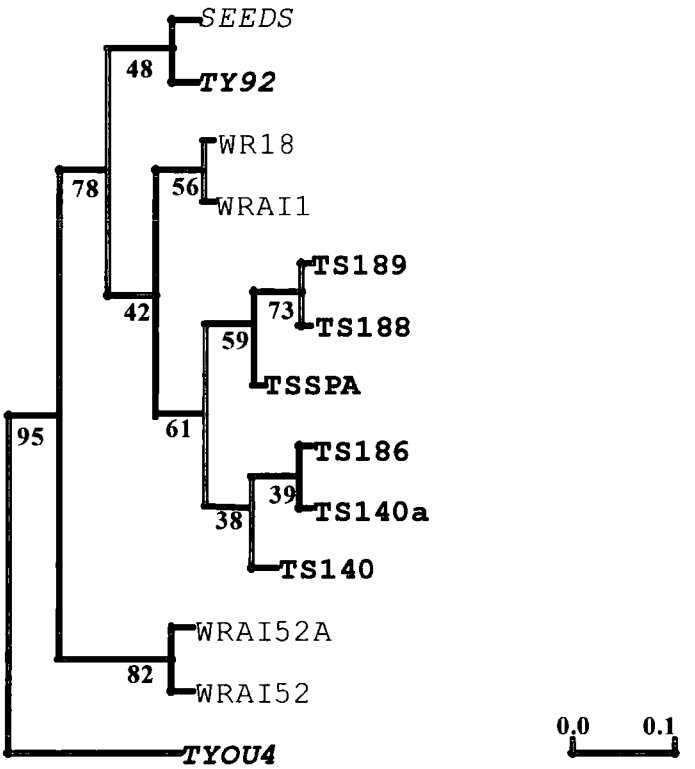
The GENDIST program was use to compute genetic distance between populations and the results were used to produce trees using the FITCH, KITCH and NEIGHBOUR programs (section 3.12.3)

Figure 5.10 Tree produced using the FITCH program. Populations from the Tees are shown in bold, those from the Tyne in bold italics and the commercial seed population is in italics. Numbers indicate the percentage of bootstrap replicates that support each node.



The tree produced using FITCH (Figure 5.10) differed from that produced using CONTML (Figure 5.9) in the arrangement of populations in the Tees and the Wear. Populations Tees Spa 3, Tees 188 and 189 were each located on their own branch, separate from all other populations. Tees 140, 140a and 186 were located along the same branch with Tees 186 and 140 were grouped together. The Wear populations were all found on the same branch, but populations Wear 18 and Wear Rainton 1 were not grouped together, as they had been in the CONTML tree (Figure 5.9).

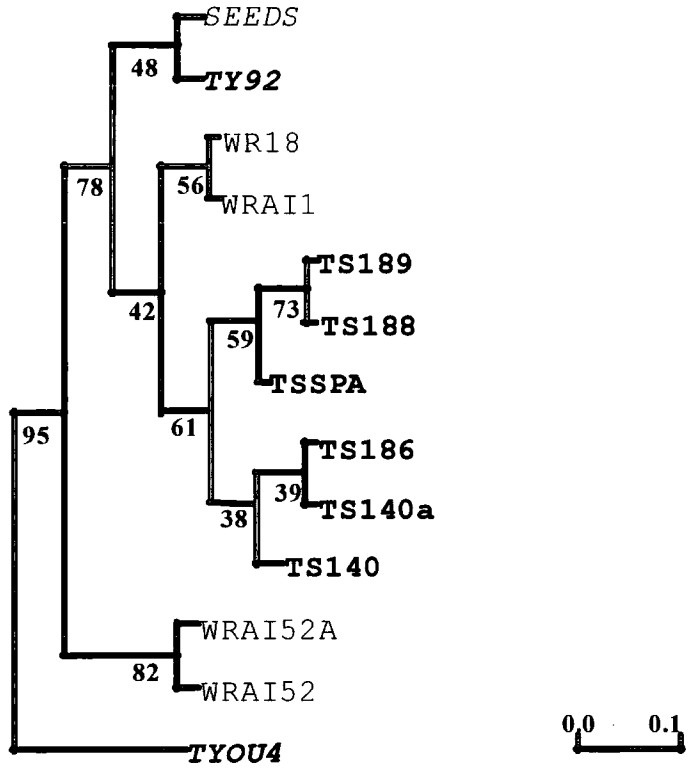
Figure 5.11 Tree produced using the KITSCH program. Populations from the Tees are shown in bold, those from the Tyne in bold italics and the commercial seed population is in italics. . Numbers indicate the percentage of bootstrap replicates that support each node.



The tree produced using KITSCH grouped all populations in the Tees together, placing population Tees 186 on the same branch as populations Tees 140 and 140a (Figure 5.11). The arrangement for populations in the Wear differed from all other methods of tree construction by splitting populations Wear Rainton 52 and 52a from the other Wear populations.

The tree produced using NEIGHBOR arranged the populations from the Tees and the Wear in the same way as CONTML, differing only by placing populations Tees 140 and Tyne Ouse 4 along a branch that was shared with the Wear populations

Figure 5.12 Tree produced using the NEIGHBOR program. Populations from the Tees are shown in bold, those from the Tyne in bold italics and the commercial seed population is in italics. Numbers indicate the percentage of bootstrap replicates that support each node.



5.5 Genetic Variation in the Tees

The values of F_{ST}, Rho_{ST} and $(\delta\mu)^2$ comparisons of population Tees 189 with population Tees 188 were very low and that of F_{ST}, not significantly different from zero (Tables 5.4-5.6). Table 5.1 shows that for all loci, the two populations were similar in both the alleles present and their relative proportions.

The values of F_{ST} comparing population Tees 186 with that of Tees 188 and Tees 189 were 0.17 and 0.19, and were significantly different from zero (Table 5.4). Genetic distance trees were produced using four different methods and none of these grouped population Tees 186 with 188 or 189 (Figures 5.9-5.12), despite their close geographic proximity.

The F_{ST} values for comparisons of populations Tees 140 and Tees 140a were small, but significantly different from zero (Table 5.4). The value of $(\delta\mu)^2$ for the comparison of population Tees 140a with population Tees 140 was larger than for the comparison between population Tees 140a and Tees Spa 3 and between population Tees 140a and Tees 188 (Table 5.6). The genotype assignment test comparing populations Tees 140 and Tees 140a (Figure 5.7d) showed there to be much greater differences than between a similar comparison of *H. mantegazzianum* (Figure 4.7d). Further evidence of the difference between populations Tees 140 and 140a was that the two populations were never grouped together in any of the four genetic distance trees produced (Figures 5.9-5.12). However, considered together, populationns Tees 140 and Tees 140a had a number of differences from all other populations in the Tees. Populations Tees 140 and 140a had seven alleles between them that were private to that location, whereas there was only one other private allele in any of the four other populations from the Tees (Table 5.1).

The values of F_{ST} comparisons between population Tees Spa 3 and other populations in the Tees were mainly quite low (Table 5.4) and populations Tees 188 and 189 were not significantly different from population Tees Spa 3. The corresponding values of Rho_{ST} were also low (Table 5.5), despite the relatively large distance between Tees Spa 3 and the other populations. This may responsible for the lack of a pattern of isolation by distance (Figure 5.8).

5.6 Genetic Variation in the Wear

The F_{ST} value for the populations Wear Rainton 52 and 52a was relatively low, but significantly different at the level of $P<0.05$ (Table 5.4). Further differences between the populations can be seen in the assignment test carried out on the two populations (Figure 5.7a), which was comparable to the differences between populations Tees 140 and 140a. In three out of the four microsatellite loci, the same alleles were present, although in different proportions, but at one locus, A2, seven alleles were present, at low numbers, in one population but not the other (Table 5.1). However, genetic distance trees produced using four different methods all grouped the two populations together (Figures 5.9-5.12) and the population were monomorphic for the same cpDNA micrsatellite allele.

F_{ST} comparisons between population Wear Rainton 1 and that of populations Wear Rainton 52 and Wear Rainton 52a were significantly different (Table 5.4), and Rho_{ST} and $(\delta\mu)^2$ values were also high (Tables 5.5 and 5.6). At locus A2, the most common allele in population Wear Rainton 1 was present only once in populations Wear Rainton 52 and was not observed in population Wear 52a (Table 5.1). There were also large differences between the alleles present in the chloroplast locus in these populations. Allele 206 was not present in the populations at the top of the tributary but was the most common allele at the bottom of the tributary (Table 5.1).

The F_{ST} comparison between populations Wear 18 and Wear Rainton 1 were not significantly different, Rho_{ST} values were zero and $(\delta\mu)^2$ values were very low (Tables 5.4-5.6). Genotype assignment tests comparing populations Wear Rainton 1 and Wear18 showed these populations to be much more similar than either pairs of populations Wear Rainton 52 and 52a or Wear Rainton 1 and 52a (Figure 5.7).

There was a significant positive relationship between geographic and genetic distance for populations in the Wear (Figure 5.9).

5.7 Genetic Variation in the Tyne

The F_{ST} value comparing the two populations in the Tyne was high (Table 5.4) and Rho_{ST} and $(\delta\mu)^2$ values were also high (Tables 5.5 and 5.6), showing there to be a very significant difference between the two populations. The very clear differences between these populations can also be seen in the assignment test in Figure 5.7.

Population Tyne 92 was most similar to the commercial seeds for F_{ST}, Rho_{ST} and $(\delta\mu)^2$ values (Tables 5.4-5.6), and genetic distance trees produced using four different methods all grouped the two populations together (Figures 5.9-5.12).

Chapter Six

Discussion

6.1 *Heracleum mantegazzianum*

A significant amount of both within and between population variation was found for *H. mantegazzianum.* This was borne out both in the number of alleles found at all nuclear genomic microsatellite loci (Figure 4.2) and in the relatively large F_{ST} and Rho_{ST} values (Tables 4.4 and 4.5). The chloroplast microsatellite locus was monomorphic for all populations except for population Tyne Ouse 4 in the Ouse tributary of the Tyne, which also held private alleles at all other microsatellite loci (Table 4.1).

6.1.1 *Heracleum mantegazzianum* in the Tees

Populations Tees 162 and Tees 162a were sampled from the same location two years apart and there was found to be very little genetic variation between them (F_{ST} and Rho_{ST} values were zero) (Tables 4.4 and 4.5). This was expected considering that the population is the second most upstream in the Tees and stems from source from a nearby garden (of which there were less than ten individuals on the riverbank). A number of alleles were found in only one year and this may reflect the temporal variation in a monocarpic species, or sampling error. In each case, roughly half the individuals present in the 500 m stretch were sampled and so the different alleles found between years may have resulted from differences in the individuals sampled. The species most commonly flowers in the third year, and since it is monocarpic, many of the individuals sampled would have died in the two years between sampling.

Population Tees 152 was sampled from 5 km downstream of populations Tees 162 and Tees 162a at a bend in the river, where it would be expected that any floating seeds from upstream may become deposited (Wadsworth *et al.* 2000). There was no significant genetic variation between population Tees 152 and those upstream of it, with values of F_{ST} and Rho_{ST} values of zero, (section 4.5).

Population Tees 129 was the next most downstream population, located 11.5 km downstream of population Tees 152 and there was greater genetic variation between these two populations than between population Tees 152 and Tees 162 (section 4.5). F_{ST} values between population Tees 129 and those upstream of it were significantly different from zero (Table 4.4). This would be expected, given that population Tees 152 is geographically more than twice as far from population Tees 129 as it is to population Tees 162. At all loci, there were no alleles found in population Tees 129 that were not also present in the upstream populations (Figure 4.5). This and the relative similarity between population Tees 129 and those upstream of it suggest that this population arose as a result of seed dispersal from upstream populations rather than from an independent introduction. Gene flow between populations is predominantly unidirectional, since seed dispersal by water occurs in a downstream direction. Differences between the most upstream populations and those further downstream will reflect the number of founders that gave rise to the more downstream populations and the level of gene flow between them as the higher the gene flow, the less differentiation there would be expected to be.

Population Tees 59 was located a further 35 km downstream from population Tees 129 and downstream of the town of Darlington. Population Tees 59 was significantly different from all other population in the Tees (section 4.5) with values of Rho_{ST} all above 0.3 (Table 4.5), which suggests that this population has arisen from an independent introduction. The first record of *H. mantegazzianum* in the Tees was at this location in 1944 (Figure 6.1), and this provides further evidence that there was an independent introduction into the Tees at this location.

The test of isolation by distance for populations in the Tees shows a significant relationship between genetic and geographic distance (Figure 4.8a). However, this positive relationship was due to populations less than 20 km apart being very similar. Populations greater than this distance apart did not show a positive correlation between geographic distance and genetic distance. This was because all population comparisons between populations greater than 20 km apart were comparisons between population Tees 59 and those upstream of it. As there is strong evidence that population Tees 59 arose from an independent introduction, the lack of a pattern of isolation by distance would reflect this. The river upstream of population Tees 59 has a number of large meanders, which would reduce the possibility of seed dispersal from upstream populations. Genetic distance trees grouped together all populations in the Tees except for population Tees 59 (Figures 4.10-4.13). Therefore, there appear to

have been at least two independent introductions into the Tees, upstream of population Tees 162 and in between populations Tees 129 and Tees 59. Population Tees 129 was similar to populations upstream of it, but not as similar as those closer together.

6.1.2 *Heracleum mantegazzianum* in the Wear

Populations Wear 71 and Wear 77 were the two most upstream populations sampled from the Wear and are both located in Durham City. The two populations were situated only 3 km apart but there was significant genetic variation between them (section 4.6), and populations much further apart in the Tees were much more similar. This suggests that the populations may have arisen from independent introductions, which would not be unexpected considering their locality. The banks of the Wear at section 71 are tall and steep, which may make deposition of floating seeds unlikely to occur.

Population Wear 46 was located 13.5 km downstream of population Wear 71 and 16.5 km downstream from population Wear 77, but was more similar to population Wear 77 (section 4.6). Assignment tests show populations Wear 46 and Wear 77 to be roughly as similar as populations Tees 129 and Tees 152 were from each other (Figure 4.7). Similar distances separate both pairs of populations, suggesting that levels of gene flow may be similar across such a distance in both the Tees and the Wear.

Population Wear 35 was significantly different from all upstream populations (section 4.6). There were a several hundred individuals in the three km downstream from population Wear 35. This suggests, as do the differences between this population and those upstream of it, that there may have been an independent introduction into the area, which includes the grounds of Lumley Castle. However, the allele may have occurred in the upstream populations at low frequencies but either was not sampled, or may have been lost by drift and similarly its high frequency in population Wear 35 may have arisen from the Founder effect. In either case, there is likely to be very low levels of gene flow between population Wear 35 and those upstream of it.

Population Wear 18 was more similar to population Wear 46 than to population Wear 35 (to which it was 4.5 km closer) with Rho_{ST} values between population Wear 18 and Wear 35 and 46 0.21 and 0.03 respectively (Table 4.5). Three out of four genetic distance trees grouped population Wear 18 with the population from London, rather than with other populations from the Wear (Figures 4.10; 4.11; 4.13). This

suggests that there was an independent introduction between population Wear 18 and Wear 35. This is very likely to have been the case since the first record of *H. mantegazzianum* in the Wear was ten km upstream of Wear 18 in 1954; six years before it was recorded in Durham city (at population Wear 77) (Biological Records Centre, Monkswood).

The result of a test for isolation by distance for populations in the Wear did not reveal any clear relationship (Figure 4.8). This may be a reflection of the greater number of introductions into the area compared with the Tees. This would be expected considering that the populations in the Tees were mostly upstream of any town, whereas populations in the Wear were all either in the centre or downstream of Durham City.

With the exception of populations Wear 46 and 77, all other population pairs situated less than 20 km apart on the Wear were quite different from one another in terms of F_{ST} and Rho_{ST} values, and genetic distance trees. This may be a reflection of a number of introductions into the Wear, and of low levels of gene flow between populations. The banks of the Wear are steeper than those in the areas sampled in the Tees and therefore, even though seeds may be dispersing long distances in the Wear, they may not reach downstream populations because they are much less likely to be washed up onto to them.

6.1.3 *Heracleum mantegazzianum* in the Tyne

The two populations sampled from the Tyne were separated 28 km. Population Tyne Ouse 4 was sampled from a tributary in the centre of Newcastle-upon-Tyne whilst population Tyne 94 was sampled from a rural area in between Hexham and Newcastle-upon-Tyne. There was significant genetic variation between these populations and population Tyne Ouse 4 had four private alleles at nuclear loci and three private alleles at the chloroplast locus (section 4.7). This suggests that population Tyne Ouse 4 arose as a result of an independent introduction that may have come from a very different source from all other populations. The relatively large number of chloroplast alleles (considering that all other populations were monomorphic for the same allele) may suggest that there has been more than one introduction at the location of Tyne Ouse 4, which is possible, given that the tributary is in the centre of Newcastle.

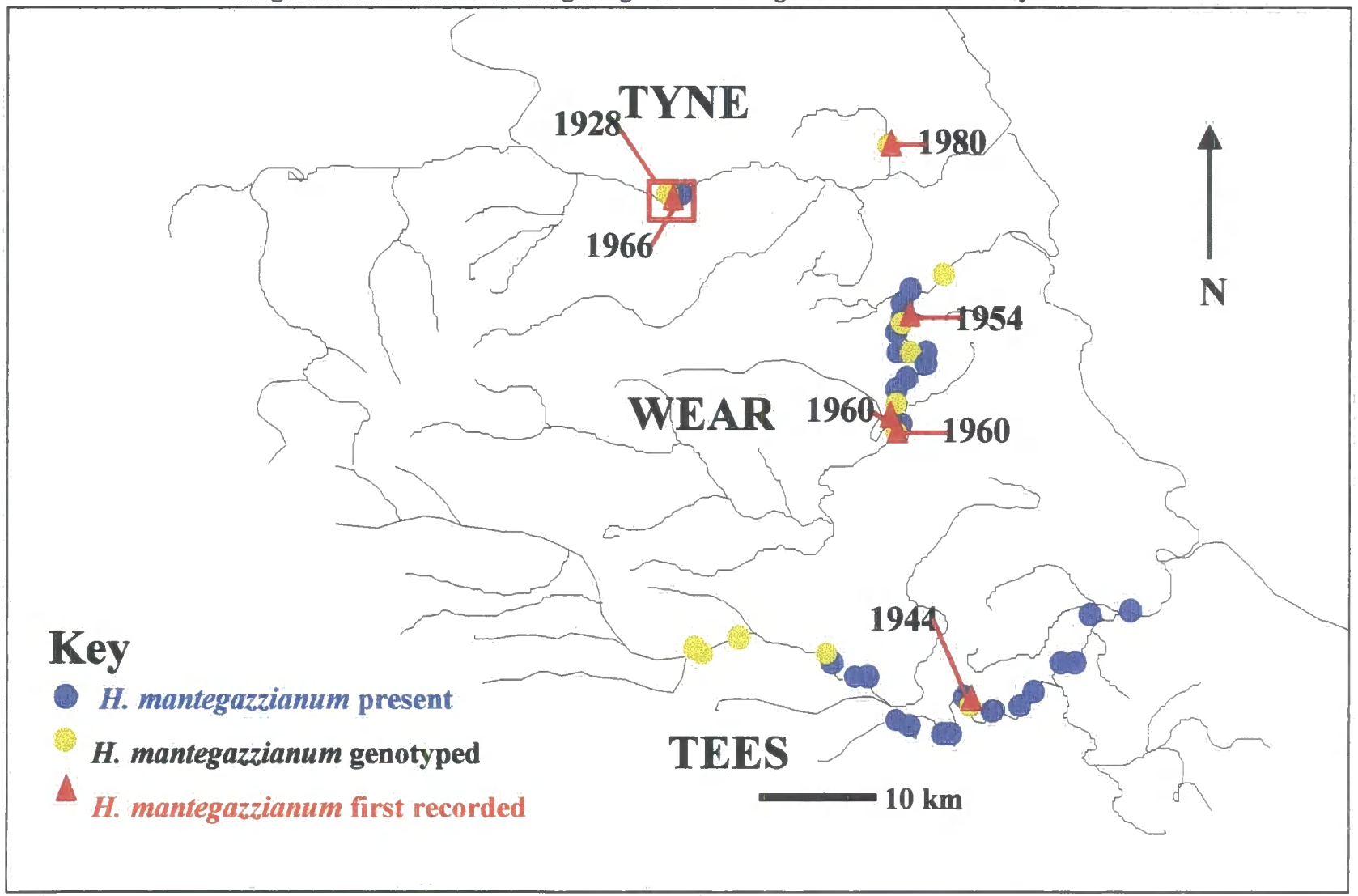
6.1.4 Catchment comparisons of *H. mantegazzianum*

The F_{ST} values of catchment comparisons were all significantly different from zero at the level of $P<0.05$ (Table 4.3). The population sampled from the north of London was different from all other populations, although F_{ST} values were not greater than for other cross-catchment population comparisons (Table 4.4). Populations within catchments were more similar than those between catchments, with the exception of population Tees 59 and the populations from the Tyne. Population Tyne 94 was more different from Tyne Ouse 4 (in terms of $(\delta\mu)^2$ and Rho_{ST} values) than any other population (Tables 4.5 and 4.6). Population Tees 59 was also found to be more similar to populations in the Wear and Tyne than to other populations in the Tees.

The F_{ST} values in Table 4.3 show that there was less variation between the Wear and London than between the Wear and the Tees. This can also be seen by comparing genotype assignment test results in Figures 4.3a and 4.3e. This suggests that introductions into each catchment came from a source outside the Northeast of England and that there has not been any mixing between catchments. This is again borne out by there being no relationship between genetic and geographic distance when considering catchments as a whole (Figure 4.4).

Sites and dates of first records are shown in Figure 6.1. Population Tees 59 was located at the site of the first record of the species in the Tees, and therefore there is likely to have been an introduction at this location. As population Tees 59 was located downstream of the town of Darlington, it would not be unexpected for there to have been at least one introduction in this area. The populations in the Wear as a whole were not as similar as those in the upper Tees (Figure 4.10), which could be because the populations in the Wear were all at or downstream of Durham city, so there is a greater likelihood of multiple introductions into the Wear than the Tees (where all but one population was upstream of Darlington). In the Wear, the two most upstream populations, situated on either side of a wide meander in the river in the city of Durham, were genetically distinct. Population Wear 46 was very similar to population Wear 77; the F_{ST} values between the two populations were not significantly different and so it appears that there was downstream dispersal between the two populations. However, at all loci in population Wear 46, there were a number of alleles that were not observed in any of the upstream populations. Therefore, there may also have been

Figure 6.1 First Recorded Sightings of *H. mantegazzianum* in the Study Area



a further introduction into the Wear at some point between populations Wear 71 and Wear 46.

6.2 *Impatiens glandulifera*

Impatiens glandulifera has a much wider distribution in the northeast of England than *H. mantegazzianum* and is present in a number of tributaries (Figure 5.1). *Impatiens glandulifera* is currently available for sale in garden centres, has been present in the study area for longer than *H. mantegazzianum* and has a greater potential for downstream dispersal (Wadsworth *et al.* 1997). These factors suggest that the pattern of variation observed would be expected to be very complex.

6.2.1 *I. glandulifera* in the Tees

Population Tees 189, at which there were hundreds of individuals, was the most upstream in the Tees and population Tees 188 (with only 35 individuals) the next downstream. There was no significant variation between these populations in terms of F_{ST} values, as would be expected (section 5.5).

Population Tees 186 was sampled from just one km downstream of population Tees 188, but was found to be significantly different from the populations upstream of it (section 5.5). Considering the great ability of *I. glandulifera* for downstream dispersal, this was unexpected. Population Tees 186 has just one allele of one microsatellite locus (locus A2) that was not found in the upstream populations (Table 5.1), although this allele may have been present in individuals not sampled upstream, considering the large number of individuals in the locality of population Tees 189. However, the proportions of alleles in population Tees 186 were different from those of populations Tees 188 and Tees 189 at all loci, especially the chloroplast locus (Table 5.1). Therefore, if population Tees 186 arose as a result of downstream dispersal, the level of gene flow would be lower than expected by considering the geographic distance between populations. The river is very wide, straight and quite fast flowing at the Tees 186 stretch, making it difficult for seeds to become deposited or establish themselves there. Therefore, despite its proximity to populations Tees 188 and Tees 189, seed dispersal from these populations to Tees 186 may be quite rare.

Populations Tees 140 and 140a were sampled from the same location two years apart. *Impatiens glandulifera* is an annual and as such, temporal variation may be expected to be greater than for populations of *H. mantegazzianum* and, as expected, the populations of *I. glandulifera* were more different from one another with F_{ST} values that were small but significantly different from zero (Table 5.4; section 5.5).

Population Tees Spa 3 was sampled from 1.5 km up a very narrow tributary of the Tees and so is highly likely to have arisen from an anthropogenic introduction. Gene flow between this and other populations would be expected to be limited to that facilitated by pollinators. F_{ST} comparisons between population Tees Spa 3 and other populations in the Tees were mainly quite low (Table 5.4) and populations Tees 188 and 189 were not significantly different from population Tees Spa 3. An explanation for this is that the populations may have been introduced from the same source, or the population could have arisen from an anthropogenic spread from the main river Tees to the tributary. The lack of differentiation of the Tees Spa 3 population (which was unexpected considering there were only 40 individuals) from the other Tees populations may be due to high levels of gene flow via pollen. This is borne out by the greater differentiation in cpDNA allele frequencies between Tees Spa 3 and the other populations in the Tees (Tables 5.2 and 5.9) than found in nuclear variation (as the chloroplast is maternally inherited).

The test of isolation by distance for populations in the Tees found no relationship between geographic and genetic distance (Figure 5.8a). Possible causes of a lack of such a relationship could be that there were a large number of introductions into the area, that levels of gene flow are very high between all populations or that being an annual with no seedbank (Grime *et al.* 1988, Beerling & Perrins 1993), populations are subject to large temporal changes (which could make isolation by distance more difficult to detect than in species with more stable numbers of indibviduals and levels of variation). The low levels of between population variation makes it unlikely for there to have been many independent introductions, and the most likely explanation for lack of a pattern of isolation by distance is that temporal variation within populations is high. This was seen in the comparison between populations Tees 140 and 140a. There appears to be less gene flow between population Tees 140/140a and those upstream of it (which are all located much closer together).

6.2.2 *I. glandulifera* in the Wear

Populations Wear Rainton 52 and Wear Rainton 52a were sampled from the same location two years apart at the top of the Rainton tributary of the Wear, but F_{ST} values and assignment tests showed there to be a significant amount of genetic variation between them (section 5.6). The total number of individuals present was estimated at about 50, and the population was taken from the top of the tributary. The differences in relative proportions of alleles may be attributed to genetic drift in this annual species. Alternatively, the population may have received pollinators that had come from downstream populations. Despite the differences between the two populations, genetic distance trees produced using four different methods all grouped the two populations together (Figures 5.9-5.12).

Population Wear Rainton 1 was sampled from the bottom of the tributary from which populations Wear Rainton 52 and 52a were taken, but population Wear Rainton 1 was adjacent to the main Wear river. There were several differences in the alleles present and significant genetic variation was found between populations Wear Rainton 1 and the two populations upstream (section 5.6). Population Wear Rainton 1 was sampled from 25.5 km downstream of populations Wear Rainton 52 and 52a and the tributary is very narrow and winding and so downstream dispersal would be expected to be very limited and upstream dispersal of seeds in effect negligible. However, pollen transfer between the populations would be possible, although probably quite limited.

Population Wear 18 was sampled nine km downstream from where the section from which population Wear Rainton 1 was sampled meets the Wear river. These populations would therefore, be likely to be similar and this was found to be the case as F_{ST} values were not significantly different from zero and Rho_{ST} and $(\delta\mu)^2$ values were also low (section 5.6). There was a positive relationship between geographic and genetic distance (Figure 5.9), although, because of the sampling strategy employed, sampling three populations up a tributary, it is difficult to make a comparison with the Tees. The observed pattern of isolation by distance in the Wear may be due to the distribution of locations at which the populations were sampled. Four populations were sampled; two were from the same location at the end of a tributary and the other two were situated close together, on the main river. Therefore, the two sets of populations are likely to have arisen from different introductions and the pattern of isolation by distance would simply reflect this.

6.2.3 *I. glandulifera* in the Tyne

Populations of *I. glandulifera* were sampled from the main river of the Tyne and from the Ouse tributary in the centre of Newcastle-upon-Tyne. The F_{ST} value comparing the two populations in the Tyne was at 0.24, high (Table 5.4), Rho_{ST} and $(\delta\mu)^2$ values were also high (Tables 5.5 and 5.6) and all three values were the highest observed for any two populations in the same catchment. These findings were not unexpected considering the large geographical distance between the populations and that population Tyne Ouse 4 was two km up a tributary.

6.2.4 Commercial seeds of *I. glandulifera*

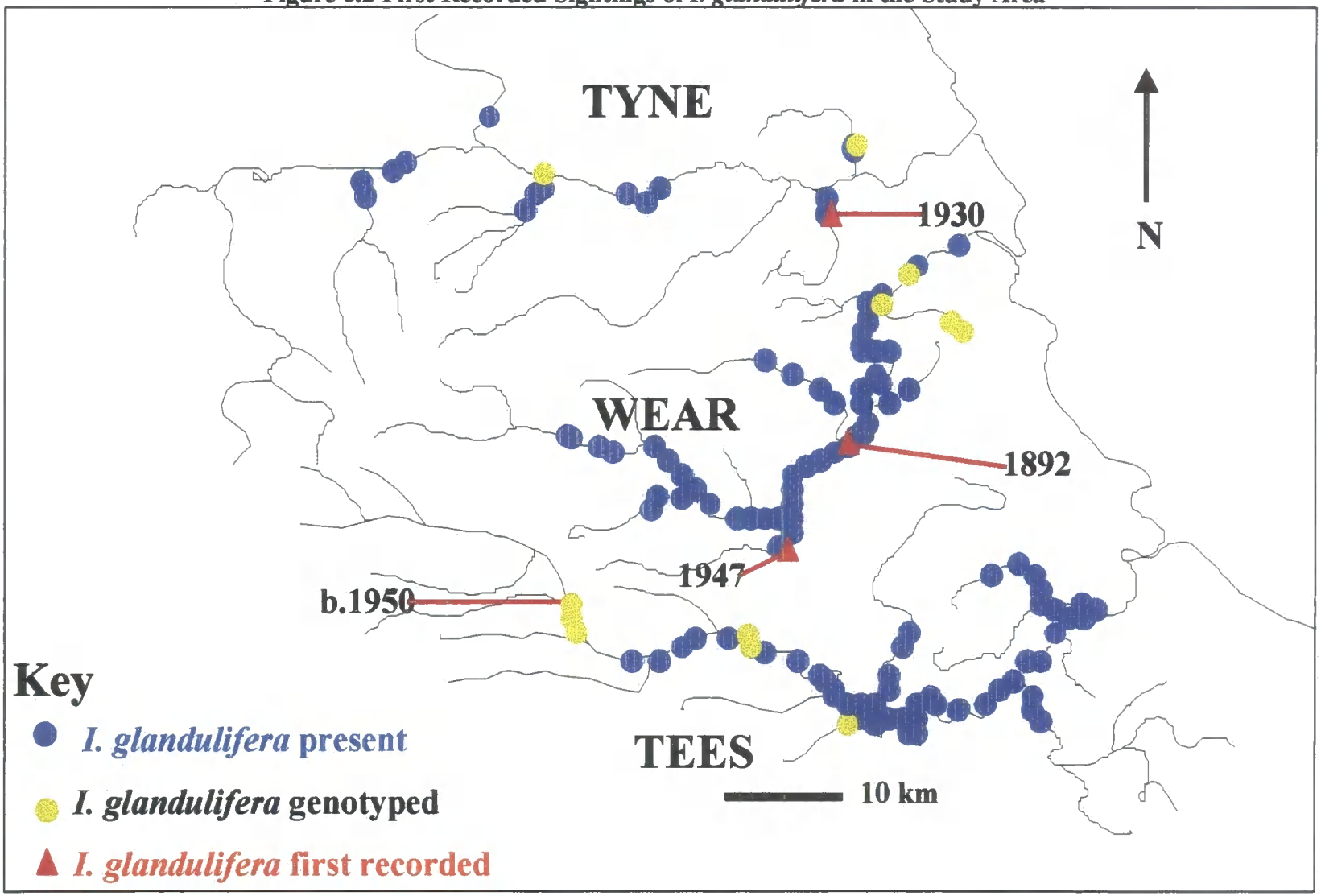
Impatiens glandulifera seeds are sold at a number of garden centres throughout Northeast England. Seeds found for sale in Durham City were genotyped in order to compare them with populations along the river, which may have come from garden escapes. The F_{ST} values comparing the commercial seeds with all other populations varied between 0.05 and 0.38, but all differences were significant at the level of $P<0.05$ (Table 5.4). The commercial seeds were most similar to population Tyne 92 for F_{ST}, Rho_{ST} and $(\delta\mu)^2$ values (Tables 5.4-5.6), and genetic distance trees produced using four different methods all grouped the two populations together (Figures 5.9-5.12). The Tyne 92 population may have grown from seeds bought from this seed company.

6.2.5 Catchment comparisons of *I. glandulifera*

The F_{ST} values comparing the Tees, Wear and Tyne catchments were all significantly different, although the average value was lower than that of the average comparison between catchments for *H. mantegazzianum* (0.12 and 0.18 respectively) (Table 5.4). Genotype assignment tests comparing the catchments also showed there to be less differentiation between catchments for *I. glandulifera* than for *H. mantegazzianum* (Figures 5.3 and 4.3). A possible explanation for this is that since *I. glandulifera* was much more commonly found in gardens, seeds could have been taken from plants in one catchment and taken to another catchment. Alternatively, seeds were commonly available at garden centres and the same source may have been introduced into more than one catchment, which is likely to have occurred, given the popularity of the

species in gardens. The sources of introductions into the different catchments could have been similar. However, the most common allele of microsatellite locus A21 in the Tees was not found at all in the Wear, suggesting that had not been any movement from material in the Tees to the Wear (Figure 5.6). The overall amount of variation in both species in Britain is dependent upon the number and sources of seeds originally introduced. There may be less overall variation in *I. glandulifera* because of its introduction history. The location of first records of *I. glandulifera* is shown in Figure 6.2. It is likely that there have been many introductions into the catchments, because of the wide availability of seeds and the large numbers of gardens observed to have been growing the species.

Figure 6.2 First Recorded Sightings of *I. glandulifera* in the Study Area



6.3 Comparisons between *H. mantegazzianum* and *I. glandulifera*

There was a greater spatial structure to the genetic variation found in *H. mantegazzianum* than in *I. glandulifera*. *Heracleum mantegazzianum* had greater between catchment variation, and within a catchment, there were greater differences between populations. The average Rho_{ST} value for within population comparisons was 0.09 for *I. glandulifera* and 0.19 for *H. mantegazzianum*. The reasons for there being less structure found in *I. glandulifera* may be that the species has been present in the study area for 36 years longer than *H. mantegazzianum*, is much more common in gardens and is sold at garden centres, has a greater dispersal ability and its pollinators have larger home ranges.

However, there were greater differences between populations less than ten km apart and between those at the same location sampled two years apart in *I. glandulifera* than in *H. mantegazzianum*. Levels of gene flow amongst populations of *H. mantegazzianum* less than ten km apart appear to be high, particularly in the Tees, where the riverbanks were very flat and the river meanders, so facilitating the deposition of seeds. There may have been greater differences between populations of *I. glandulifera* because a smaller proportion of each population of *I. glandulifera* was genotyped (as populations were much larger than that of *H. mantegazzianum*) and being an annual, populations of *I. glandulifera* can vary greatly from year to year. In addition, as *I. glandulifera* is thought to have no seedbank (Grime *et al.* 1988, Beerling & Perrins 1993) populations may go through frequent bottlenecks particularly as many populations are subject to winter flooding events. This would lead to increased genetic differences between populations and may explain the differences between populations of *I. glandulifera* sampled from the same location two years apart.

The different ecological characteristics of the species and differences in their introduction into the area are shown in Table 6.1. Populations of *I. glandulifera* are more prone to extinction than that of *H. mantegazzianum* because *I. glandulifera* is an annual with no seedbank (Grime *et al.* 1988, Beerling & Perrins 1993). Populations of *H. mantegazzianum* that were sampled may be less ephemeral because the species does have a seedbank. Seeds have been reported to remain viable for several years (Vogt Anderson & Calor 1996), although knowledge of the seedbank is incomplete.

However, populations were only chosen to be genotyped where there were more than twenty-four individuals present. There were a number of populations consisting of

Table 6.1. Comparisons of ecological and historical features between *H. mantegazzianum* and *I. glandulifera*

Feature	***H. mantegazzianum***	***I. glandulifera***
Distribution	Widespread but sparse	Widespread
Life history	Monocarpic perennial	Annual
Average population size	~30	~200
Self-pollinaton ability	Can self-pollinate	Can self-pollinate
No. of seeds per plant	c. 10,000	c. 800
Seedbank	Seeds viable for 7-15 years	No viable seedbank
Pollinators	Diptera and hymenoptera	*Bombus lucorum*
Pollinator range size (km^2)	1	5
Maximum dispersal distance (km) (Wadsworth *et al.* 2000)	10	20
Date of introduction into study area	1920s	1892
Introduction history	Introduced into large estates	Available at garden centres and popular in gardens

five or fewer individuals, and these are likely to be short-lived. Each plant can produce several thousand seeds, many of which fall very close to the parent plant. However only a proportion of these seeds will germinate and year old seedlings have been found to suffer very high mortality rates, which can result in the initial large numbers of seedlings becoming completely eliminated (Willis 1999).

Impatiens glandulifera has a wider distribution than *H. mantegazzianum* in all three river catchments, being present further upstream in each catchment, in more of the 500 m River Corridor Survey sections and in many more tributaries. This suggests that not only has *I. glandulifera* been introduced into a greater number of locations, but also that it has been able to spread into a greater number of suitable areas and its range is not thought to be expanding (Willis 1999). Willis (1999) tested the ecological constraints of the two species by planting the species in areas outside their current distribution and following their progress over three years. The upstream distribution of *I. glandulifera* was restricted by low seed production. In contrast, *H. mantegazzianum*

was found to have not yet to reached its full potential distribution and is limited by lack of introduction into more upstream areas and also into tributaries. Dispersal upstream by natural means would occur very slowly.

National surveys of the distribution of the two species carried out by the Environment Agency shows the presence of *I. glandulifera* in many more 10 X 10 km map squares than *H. mantegazzianum* (Dawson & Holland 1999). This study identified a number of methods of anthropogenic spread of both species such as clearance of spoil from sites of development, vegetation maintenance of riverbanks, and transfer of seeds from footware and tyres. All these methods would be more likely to result in the movement of seeds within a catchment, although cross catchment movement is possible.

When considering the amount of gene flow in the two species, comparisons are difficult because a smaller total area of each catchment was analysed for genetic variation in *I. glandulifera*. Additionally, as populations of *I. glandulifera* tend to consist of many more individuals than *H. mantegazzianum* and because *I. glandulifera* is an annual, for each population sampled, a smaller proportion of the total variation present in the population was sampled and therefore the estimate of gene flow is less accurate. Therefore, if populations downstream contain alleles not found in an upstream population, the chances of this being due to a lack of sampling the allele upstream, rather than to there having been an independent introduction is greater in *I. glandulifera*. Further sampling in both the Tees and the Wear would be required in order to assess this.

6.3.1 Deviations from Hardy-Weinberg equilibrium

At all nuclear microsatellite loci, both species were found to have at least one population that had a significant deficiency of heterozygotes from Hardy-Weinberg proportions. A deficiency of heterozygotes may result from non-random mating, the Wahlund effect or the presence of null alleles (Robertson & Hill 1984). Both species are self-compatible, although the levels of inbreeding were not known.

All but three populations (Wear 71, Wear 77 and London) of *H. mantegazzianum* were composed of sixty or fewer individuals, and many populations were quite isolated. These factors would increase the likelihood of non-random mating. Individual plants are often found several metres away from one another and many of the pollinators (a variety of Diptera and Hymnenoptera) tend to travel quite

short distances (Tiley *et al.* 1996). Each plant is composed of a number of umbels, and many pollinators would be expected to travel from umbel to umbel on an individual plant before moving onto a new individual. Therefore, the chances of self-fertilisation are high, which may explain the large number of observed deficiencies of heterozygotes. Populations at which there were significant deviations in two or three loci of *H. mantegazzianum* include Tees 152, Tees 162 and Wear 18, which were all relatively isolated populations of 60 or fewer individuals which were spread out over the 500m reach of each river. In these populations, the observed the deficiencies in heterozygotes could be due to the Wahlund effect.

Populations of *I. glandulifera* are generally a lot larger than those of *H. mantegazzianum* and individual plants are found much closer together. The level of non-random mating would be expected to be lower than in *H. mantegazzianum* as the pollinators are more likely to move between individuals. The proportion of loci at all populations that had a significant deficiency of heterozygotes for *H. mantegazzianum* was 0.35, whilst the corresponding proportion for *I. glandulifera* was 0.17 (not including the population consisting of commercial seeds). This shows that for *I. glandulifera*, there were fewer deviations but there was still deficiency of heterozygotes present in a number of populations. Populations of *I. glandulifera* are ephemeral and winter flooding can destroy a population, which may be recolonised by a small number of individuals. Where population numbers are low, the chances of non-random mating may be higher, and this could result in there being an observed deficiency of heterozygotes. This scenario does not explain the large numbers of alleles observed, but these could come about from subsequent dispersal from other upstream populations, though in small enough numbers to introduce new alleles but not to affect the overall proportions of heterozygotes. Individuals arising from seed dispersal from upstream populations may be homozygous for alleles which are very common in the population from which they originated, but rare in the downstream population, so adding to the deficiency in heterozygotes.

Turner *et al.* (1982) found that even in outcrossing species, where there are high levels of pollination between one plant and its nearest neighbour, this would lead to non-random mating within small groups and lower than expected levels of heterozygosity. Such a scenario is more likely to have an effect in populations comprising several different groups consisting of small numbers of plants, which is more likely to be the case with *H. mantegazzianum*, and may explain its higher proportion of homozygote excess. The bees that pollinate *I. glandulifera* have large

home ranges, but when they find a cluster of individuals, they may be likely to visit several of the individuals within a group.

Levels of non-random mating in self-compatible species vary depending on population structure (Wang 1997). The population of *I. glandulifera* in the most upstream stretch of the Tees, (Tees 189) consists of a large number of individuals in one almost continuous clump and was the most upstream at which the species occurs, so would not receive any seeds by dispersal. Population Tees 189 would therefore not be expected to be subject to the Wahlund effect, inbreeding would be expected to be low and there would not be any immigrants. Therefore, there the population would not be expected to have a deficiency of heterozygotes and, as can be seen from Table 5.7, there was no significant deviation from Hardy-Weinberg proportions.

A final possibility is that of the presence of null alleles. However, deviations from the Hardy-Weinberg equilibrium were observed in all loci of both species and given the likelihood of non-random mating, the Wahlund effect and the influence of recent migrants on levels of heterozygosity, null alleles are unlikely to have been present. In addition, no individuals sampled failed to amplify, which provides further evidence against the presence of null alleles.

6.4 Modelling the Spread of Colonising Species

Invasive species which have yet to reach the limits of their distribution provide the opportunity to study their rate of spread and the areas to which they spread. Studies of the proportion of invasive species in different habitats have found riparian habitats to be amongst the most vulnerable to invasion (Hood & Naiman 2000). The spread of riparian species with a linear dispersal via water may be easier to predict than the spread of species with other dispersal mechanisms.

6.4.1 Modelling the Spread of *H. mantegazzianum* and *I. glandulifera*

The spread of *H. mantegazzianumi* and *I. glandulifera* in riparian habitats was used to understand how environmental and ecological factors act at different spatial scales to determine their distribution by identifying which factors best correlate with the distribution of these species at different spatial scales (Collingham *et al.* 2000). Spatial

autocorrelation analysis was carried out to investigate the effect that the presence of a species in one 10 X 10 km or 2 X 2 km square has on its presence in a neighbouring square. Invasive species would be expected to show spatial autocorrelation as they spread across a landscape. Such a relationship was found for *I. glandulifera* but not for *H. mantegazzianum*. The lack of such a relationship for *H. mantegazzianum* was attributed to the importance of long-distance dispersal events, which may have been caused by humans, to the distribution of the species.

In order to investigate the spread of the species, the cell-based model MIGRATE was developed (Wadsworth *et al.* 2000). Demographic, environmental and geological parameters were used to simulate the spread of a species across a heterogenous landscape and the size and flow of rivers and tributaries were included. The riparian environment was divided into cells, with variables of the carrying capacity of each species, invasibility and persistence. The cells were divided into landcover classes and the carrying capacity of each landcover class was determined from field experiments and from the Environment Agency River Corridor Survey. A stepwise logistic regression of the distribution of the species for a number of different environmental variables (see above) was used to assess the suitability of the habitat within each cell although the maximum occupancy of a cell was taken as 1 % (Wadsworth *et al.* 2000). The model allows species to be introduced at sites where records show introductions to have taken place and the spread is determined by the characteristics of the species and the landscape from cell to cell. There was a lack of data for records of introductions, and any information on dates and locations of introductions could have improved the models (Collingham *et al.* 1997). In addition, propagule dispersal along river courses was added. Floating propagules, such as the seeds of *H. mantegazzianum* tend to become deposited at bends in the river or at floods limits well above the water surface. It was found to be much more difficult to predict the movement of seeds of *I. glandulifera*, which are transported whilst submerged and can germinate under water (Wadsworth *et al.* 1997).

The model was used to predict the distribution of the species in the Wear river catchment using a cell size of 500 m by 500 m and predicts the distribution of the species after each time step, which was one year. The model was found to more accurately predict the spread of *I. glandulifera*, with more than half of the presences in cells being correctly predicted and 94 % of absences were correctly predicted. The model was less successful at simulating the spread of *H. mantegazzianum* and different simulations gave more variable results for predicting its occurrence

(Wadsworth *et al.* 2000). The poor predicting ability for *H. mantegazzianum* was thought to be at least partly due to the lack of knowledge about the number and location of anthropogenic introductions of the species into the catchment. This lack of knowledge would have been aided by using genetic variation to predict likely introductions.

The simulation was expanded to the scale of the whole of England and Wales, with the distribution in 1960 predicted from the known distribution in 1940. The model was not in this case as good a predictor as when applied to just the Wear and the discrepancy was thought to be due to the difficulty in estimating the carrying capacity of cells. The simulation was rerun just for *I. glandulifera* over the same time period but the cells which actually contained *I. glandulifera* in 1960 were taken to be the only cells into which the species could spread. With this scenario, 73 % of presences were predicted, although over a quarter of actual presences were not found by MIGRATE to be able to have been colonised. The poor ability of the model to determine the spread of the species at larger scales was attributed to both lower resolution of environmental data and to a difficulty in predicting long-distance dispersal events. These events have a greater impact at larger spatial scales. The importance of long-distance dispersal can be assessed by analysing genetic variation and it was in light of the usefulness of the study of genetic variation that this current study was undertaken. The resulting genetic variation found, which predicted that there were a number of independent introductions into each catchment for both species, helped to justify the finding that long-distance dispersal events were responsible for the lack of accuracy of the model at larger spatial scales.

6.4.2 Predicting the Future Spread of *H. mantegazzianum*

Ecological studies and the short time that *H. mantegazzianum* has been present in Northeast England indicate that the species is still expanding its range in all river catchments. Analysis of genetic variation provided evidence that the species has spread at least 15 km in both the Tees and the Wear river catchments although the species may have already spread considerably further. Efforts to control the spread of the species currently consist of using herbicide to destroy flowering individuals in the centre of Durham City and in the grounds of Lumley Castle and surrounding golf course in the Wear. In the Tees, control efforts are directed at the most downstream populations, but this will not prevent recolonisation from upstream populations. If

control efforts are not increased, eventually all suitable habitats will be susceptible to invasion. However, spread upstream and into tributaries would occur very slowly by dispersal of seeds by wind and may depend upon an anthropogenic source of dispersal. It is illegal to introduce *H. mantegazzianum* into natural habitats because of the danger from its poison but introduction into new areas could occur by any of the ways given in section 6.3. These potential sources of spread would be most likely to lead to the spread of the species within a catchment. The large size and potential health hazard of *H. mantegazzianum* currently tend to confine the plant to large gardens and estates. Therefore, introduction of the species into new catchments would be expected to be relatively rare. However, the self-compatible nature of the plant, its ability to produce a large number of seeds and disperse downstream makes it possible for just one introduction of a small number of seeds to lead to the colonisation of large stretches of a catchment.

6.4.3 Predicting the Future Spread of *I. glandulifera*

The range of *I. glandulifera* in the Northeast of England was found to have been limited in terms of upstream distribution by ecological constraints (Willis 1999). However, the species still has the potential to colonise areas within its current distribution in the gaps in its linear distribution along rivers. Seeds of the species are currently available from garden centres in Northeast England and the species was observed in a number of gardens bordering rivers in the study area. The popularity of the plant in gardens and its easy availability make the likelihood of introduction into river catchments not currently colonised greater than for *H. mantegazzianum*. In the Northeast of England, the species is not currently being controlled and therefore, the species is likely to be limited in its distribution only by time and climatic conditions.

6.4.4 Modelling the Colonisation of Other Species

Models of the spread of *H. mantegazzianum* and *I. glandulifera* stressed the importance of long-distance dispersal events in their distribution. Computer simulations of the effect of different forms of dispersal on genetic structure of a species undergoing range expansion corroborated this finding (Ibrahim *et al.*1996, Higgins & Richardson 1999). Ibrahim *et al.* (1996) carried out simulations of range expansion with stepping-stone, normal and leptokurtic modes of dispersal and the

resulting differences were examined. The aim of the study was to examine the effect of mode of colonisation, and the number of colonising individuals on population differentiation during range expansion. As a species spreads into a new area, long-distance dispersal events were found to lead to the establishment of small demes far ahead of the rest of the species distribution in the case of normal and leptokurtic dispersal patterns. These demes had low levels of genetic variation and high levels of inbreeding. The different founding demes were highly differentiated, as would be expected if they arose from different founding individuals and following genetic drift. The number of long-distance dispersal events was largest where dispersal followed a leptokurtic pattern. This resulted in the greatest number of colonising demes and the highest levels of differentiation between such demes. The colonising demes themselves had the lowest levels of variation of the three forms of dispersal and they could spread and give rise to patches which were genetically very similar. This resulting pattern of genetic variation was found to persist for hundreds of generations. Therefore, if estimates of current gene flow are made from these patterns of variation due to founding events, they may be misleading. Strand *et al.* (1996) also noted the difficulty in distinguishing between the effects of historical distribution and ongoing gene flow because they can give rise to similar patterns. This problem was encountered with *H. mantegazzianum* in the Wear and *I. glandulifera* in the Tees, as in both cases, populations close together were markedly different in terms of genetic variation. This could reflect either low levels of gene flow, resulting in genetic drift, or independent introductions. It is possible to distinguish between the two by either sampling individuals located in between the two populations or by resampling the populations in later years. If the differences are due to historical introductions, current gene flow levels may be much higher than if the populations arose from the same introduction and became differentiated over time. Therefore, resampling years later, if there is less variation between populations, the differences are likely to be due to historical introductions.

Higgins *et al.* (1996) modelled the pattern of colonisation of an alien pine in the South African fynbos and examined the relative importance of a range of demographic, environmental and ecological factors. Reaction-diffusion models are the most common models used to predict the spread of invasive animal species. These models assume that dispersal is independent of environmental variability and that stochastic events do not play an important part in determining the course and speed of an invasion. These assumptions may limit the ability of reaction-diffusion models to

predict the spread of plants and therefore, this model was compared with a spatially explicit, individual-based simulation model. Dispersal ability proved to be the most important of all factors tested. Other factors having an effect on the spread of the species included fecundity and frequency of fire. The inability of the reaction-diffusion model to take such factors into account limited the accuracy of its predictions and so the individual-based simulation model proved much more useful as a predictive tool.

Higgins and Richardson (1999) pointed to the occurrence of long-distance dispersal events as being responsible for the underestimation of many models in predicting the rates of plant migration. A number of different solutions to this problem have been attempted, including the incorporation of rare long-distance dispersal events into reaction-diffusion models. Higgins and Richardson (1999) approached the problem by integrating into models of spatial dispersal, statistical models which could take into account different dispersal processes. They found that not only were long-distance dispersal events important, but their interplay with other factors such as life-history features can also play a part and should also be considered. They recommend that for studies of dispersal, there should be a much greater emphasis placed on sampling rare events. This can often prove difficult in ecological studies and may be an area in which the use of genetic methods could become particularly useful. Another indication of long-distance dispersal events having played an important role in the distribution of a species is a patchy pattern of distribution. Both *H. mantegazzianum* and *I. glandulifera* had patchy distributions, although this could be due to either long-distance dispersal events or a number of anthropogenic introductions.

6.4.5 Conclusions about the Spread of Introduced Species

Heracleum mantegazzianum and *I. glandulifera* were introduced from distinct regions of the world at different times and also differed greatly in their popularity as garden plants. However, there was found to be a surprisingly large amount of genetic variation in both species within all river catchments tested. The species are likely to have been introduced from a number of independent sources into each catchment and therefore the extent and importance of human-mediated dispersal of both species may be very large. The number of introductions of the species into Britain is not known and neither is there information on how many individuals were present in each introduction. Therefore, where there is genetic evidence of independent introductions

into an area, the feasibility of there having been introductions from entirely separate sources could not be known.

The evidence for the great importance of humans in the spread of both species suggests that this is likely to be the case for many other introduced species.

6.5 Estimating Gene Flow in Plants

Gene flow estimates may be obtained using the relationship $F_{ST} = 1/(1 + 4Nm)$ as described in section 2.3.1. However, this relationship depends upon a number of assumptions, which are often violated. These assumptions are that there is no selection, no mutation, random mating within populations, that all populations contribute equally to the migrants, population sizes are constant, migration occurs at random and that there is an equilibrium between extinction and migration (Whitlock & McCauley 1999).

In the case where local extinctions and colonisations are common, such as with the two study species, especially *I. glandulifera*, population sizes will obviously vary greatly and local populations can arise from very different numbers of individuals. F_{ST} values are likely to vary temporally. Similarly, migration will not occur at random as there is a linear structure to seed dispersal and barriers to dispersal are present, such as the difficulty in crossing catchments.

Neither species has yet reached its maximum range in each catchment (Willis 1999), and it is likely that many populations have yet to reach an equilibrium between migration and genetic drift. The violations make any estimates of gene flow from F_{ST} tentative but are likely to be correct within a few orders of magnitude (Whitlock & McCauley 1999).

Gaggiotti *et al.* (1999) also found that F_{ST} based estimates of gene flow may be inaccurate, particularly when populations are composed of fewer than 500 individuals (as was often the case) and migration rates are low, because of violations of the assumptions given above. Analysis of isolation by distance may also be misleading because when based on pairwise comparisons of all populations, one population that does not fit the pattern may have a large effect and can disguise an otherwise significant pattern (Bossart & Prowell 1998). This may have had an effect on the observed lack of a pattern of isolation by distance of *I. glandulifera* populations in the

Tees. Population Tees 186 was very different from populations less than two km upstream of it. This led to populations Tees 188 and 189 being more similar to populations Tees 140 and Tees Spa 3, which were geographically much more distant.

A pattern of isolation by distance has been observed in a number of studies of natural plant populations. An example was a study of genetic variation in *Brassica oleracea* L. (wild cabbage) in Dorset using microsatellites and isozymes. A large amount of variation between populations was also found (Raybould *et al.* 1999). Isolation by distance is expected where all populations stem from the same source population or populations, and where there are moderate levels of gene flow. If levels are very high, there will be no differentiation between any of the populations and if levels are very low (or zero), all populations will be equally distinct.

6.5.1 Relative rates of seed and pollen migration

In most species of angiosperms, chloroplast inheritance occurs through the maternal lineage and so chloroplast markers can be used to determine the movement of seeds but not pollen (Ennos 1994). Gene flow for nuclear markers can occur in both and therefore higher levels of variation are often expected in nuclear markers compared to chloroplast markers (although the chloroplast markers may be more sensitive to demographic fluctuations as the effective population size is lower since they are haploid) (Comes & Abbott 1998). In addition, the mutation rate of chloroplast DNA is lower than that of nuclear DNA and this would further confound the lower variation found in chloroplast DNA (cpDNA). Ennos (1994) showed that by comparison of F_{ST} values for nuclear and chloroplast markers, the relative levels of pollen and seed dispersal can be estimated. The model for this, however, has a number of assumptions, such as populations being in equilibrium with a constant level of gene flow, which were violated in the case of colonising species. Hu and Ennos (1997) found that population differentiation was greater in maternally inherited cpDNA than in nuclear markers under the isolation-by-distance model as well as the stepping stone and island models. The cpDNA microsatellite loci of *H. mantegazzianum* and *I. glandulifera* had many fewer alleles than the nuclear markers. As seed dispersal for *H. mantegazzianum* and *I. glandulifera* was expected to be much greater than that of pollen movement, the lower levels of cpDNA variation observed may have been a result of the lower mutation rate of cpDNA.

Comparisons of relative levels of nuclear and cpDNA variation was used by Comes and Abbott (1998) to determine the relative importance of historical events (long-distance seed dispersal from glacial refugia) and gene flow. They studied a species of ragwort, whose seed dispersal potential is much wider than pollen flow, as was the case with *H. mantegazzianum* and *I. glandulifera*. They showed that recent founder events could be identified by the relative levels of variation in allozymes and cpDNA because the lower effective population size of cpDNA would lead to a greater loss in variation than in allozymes. This loss was identified by comparing levels of variation between populations in newly colonised and longer-standing regions.

6.6 Seed dispersal and colonisation

Gene flow by way of seed dispersal can be measured directly by marking seeds and establishing how far they spread. However, these studies may fail to pick up rare long-distance dispersal events, which may have large consequences in terms of genetic differentiation between distant populations (Nathan & Muller-Landau 2000). In the case of *I. glandulifera* and *H. mantegazzianum*, seeds which reach a water course are capable of dispersing several kilometres, and so obtaining a direct measure of the distance travelled by seeds or the proportion of seeds which travel a significant distance would be prohibitively difficult. The difficulty with direct measures arises because extremely large areas of river would have to be searched and very long distance dispersal events may be very rare but have important genetic consequences. Even if direct measures were made by marking and identifying seeds, many of the dispersed seeds would either land in an unsuitable area or may not germinate, so would not be relevant in terms of gene flow. The use of genetic markers overrides such problems with direct measures as only dispersed seeds that germinate are included in subsequent estimates of gene flow.

6.7 Genetic variation in colonising species

Genetic variation in a species colonising a new area would be expected to be low due to founder events and because many invasive species of plants are self-compatible. Therefore, the level of intra and inter-population genetic variation found in both study species was surprisingly high.

Studies of genetic variation of plants in their native area have often shown there to be similarly high levels of variation in an area of comparable size to that of this study. However, were the same species to colonise a new area, genetic variation would be expected to be lower (Amsellem *et al.* 2000). Genetic variation in the invasive weed *Rubus alceifolius* was measured using AFLP markers. Populations were sampled from both the species' native range and on islands to which it had been introduced. There were high levels of variation in populations in Asia, to which the species is native, but variation in the Indian Ocean islands and in Queensland, where the species had been introduced, was much lower. Lower levels of variation in introduced areas compared with native ranges have been found in a number of other comparitive studies of colonising species (Neuffer & Hurka, 1993; Novak & Mack 1993; Bjorkland & Baker 1996; Young & Murray 2000). These studies measured allozyme variation and found there to be more alleles and a greater number of polymorphic loci in native populations. Neuffer and Hurka (1993) compared allozyme variation in the weed *Capsella bursa-pastoris* from a wide range of introduced locations in North America with native populations throughout Europe. The number of genotypes found in the introduced locations was roughly half that of native locations. Similarities in genotypes between native and introduced locations enabled some North American populations to be traced back to the native region from which they arose, and led to the conclusion that there had been at least twenty independent introductions. Pérez de la Vega *et al.* (1991) was also able to find great similarities between population of *Avena barbata* in California with those in its native location in Spain. Milne and Abbott (2000) were able to identify Spain and Portugal as the areas of source material for the introduction of Rhododendron in the British Isles using cpDNA and rDNA restriction fragment length polymorphisms. Similar comparsions with native populations would have been useful for *H. mantegazzianum* and *I.glandulifera,* although many native populations covering a wide area would have had to have been sampled.

Novak and Mack (1993) found allozyme diversity in an introduced grass to be greater overall in native Eurasian areas than introduced North American regions. However, there was more within-population variation in introduced areas, which was attributed to the effects of multiple introductions into the same area. This study highlights the importance of the scale at which variation is examined. The scale over which variation in *H. mantegazzianum* and *I. glandulifera* was examined was smaller than that of most other studies. Therefore relative comparisons between populations and regions (or in this case, catchments) could give different results from those of the above studies.

Self-compatible species often have higher levels of inbreeding in introduced areas compared with their native range, since founding populations can be composed of few or even just one individual (Novak & Welfley 1997). The observed high levels of F_{IS} in populations of *H. mantegazzianum* and *I. glandulifera* (Tables 4.7 and 5.7) suggests that inbreeding occurs in many populations. Comparisons of heterozygosity levels in native populations would enable greater levels of inbreeding in the introduced areas to be detected.

Studies of variation in species whose range expanded in Europe following post-glacial expansion may provide useful comparisons with patterns of variation in invasive species. Patterns of present day genetic variation have been used to infer post-glacial colonisation patterns in a number of European species of plants and animals. Analysis of mitochondrial control region sequences in the European great tit (*Parus major major*) suggested that the species has expanded after the last ice age from a single refuge situated in the Mediterranean region (Kvist *et al.* 1999). Allozyme variation in Scottish and European populations of heather has been used to assess the effect of glaciation on the species (Mahy *et al.* 1999). The species has a continuous distribution in Scotland and there was little between population variation for populations in Scotland compared within between population variation for European populations. This was not expected considering that the largest deme of plants in western Europe occurs in Scotland because there would be expected to be more variation where there was a greater number of individuls. The low level of between population variation was attributed to the loss of variation during glaciation in Scotland, with the species having gone through a narrow bottleneck. Alexandrino *et al.* (2000) used allozymes and mitochondrial markers to determine the effect of glacial refugia followed by post-glacial colonisation on the current pattern of variation in the golden-striped salamander. The species was chosen as one in which the post-glacial

colonisation pattern may be more easily elucidated from patterns of current variation because the species is confined to areas close to streams and cannot move between suitable habitat patches. The pattern of variation observed suggested that during glaciation, there were at least two separate refuge populations, north and south of the Mondego river. The southern population appeared to have been smaller and north of the river, there may have been several populations which were able to mix following post-glacial recolonization of the area north of the river.

These studies show that much of the current pattern of variation in a number of species is attributable to events that occurred thousands of years ago. Therefore, for species which have only colonised an area in a matter of decades, as is the case with *H. mantegazzianum* and *I. glandulifera*, the number and location of introductions should be of the utmost importance in determining current genetic variation.

6.8 Population genetic structure

Many studies of variation in natural plant populations using sensitive markers such as microsatellites and AFLPs have found there to be a large amount of both within and between population variation.

In a study of the population genetic structure of self-compatible seagrass, *Zostera marina*, using microsatellites, the amount of within-population variation varied greatly between populations (Reusch *et al.* 2000). An excess of homozygotes was found in only one out of twelve populations, indicating that levels of self-fertilisation were relatively low. Populations were sampled from Europe and Canada. The European populations showed a positive relationship between genetic and geographic distance. The Canadian populations did not adhere to this pattern, as they were more similar to Northern European populations than to the less distant Baltic populations. An explanation given for this was that following glaciation, Northern Europe and Canada may have been recolonised from a common refugial population. At a geographic range similar to that at which substantial population differentiation was found in *H. mantegazzianum* and *I. glandulifera*, there was little differentiation between populations of *Z. marina*, which possibly reflects a larger level of gene flow at this distance in *Z. marina*.

A study of the population genetic structure of *Juniperus communis* using AFLPs found there to be high levels of genetic variation within populations (Van der Merwe *et al.* 2000). Some individuals in the same population were found to share only 50 % of the same bands, although there was clear structuring between populations. There was as much variation in small populations, with as few as eleven individuals, as in very large populations, possibly indicating that relatively recent habitat fragmentation had not led to a loss of variation.

6.9 Conclusions

The aim of this project was to investigate the relative effects of anthropogenic introduction, dispersal and life-history strategies on genetic variation in invasive plants. Four hypotheses were stated and these will be considered in turn.

Hypothesis 1) Populations in the same catchment are more similar than populations in different catchments.

The average F_{ST} and Rho_{ST} values for between catchment population comparisons were much greater than within catchment comparisons for both species. However there were exceptions in both species. For *H. mantegazzianum,* these were populations Tyne Ouse 4 and Tees 59 and for *I.glandulifera* these were populations Tees 186 and the populations in the Tyne. The populations of *I. glandulifera* from the Tyne were each more similar to populations in the Tees and the Wear than to each other (section 6.2). Genetic distance trees grouped populations from different catchments together for both species. Therefore, the hypothesis was not held to be correct for all populations.

Hypothesis 2) Within a catchment, there is a pattern of isolation by distance. Therefore, the pattern of variation is a reflection of the dispersal of the species.

As only two populations were sampled from the Tyne, this hypothesis was tested in only the Tees and Wear. There was a positive correlation between genetic and

geographic distance for *H. mantegazzianum* in the Tees and for *I. glandulifera* in the Wear. However when populations at a minimum of 20 km apart are considered, there is no correlation between geographic and genetic distance (Figures 4.8a and 5.8b). This may be due to the effects of multiple introductions.

Populations of *H. mantegazzianum* in the Wear did not show a pattern of isolation by distance and this is likely to be because there is evidence of at least two introductions into the Wear in the sampling area. Populations of *I. glandulifera* in the Tees did not show a pattern of isolation by distance. There is no evidence for multiple introductions in this location, and the lack of a relationship may be due to high levels of temporal variation. This would not have had such an effect in the Wear largely because of the sampling strategy employed (section 6.2.2).

Of the three factors whose effects on variation were examined, when considering isolation by distance, it appears that life-history strategy most affects *I. glandulifera*, whereas human introduction was most important for *H. mantegazzianum.*

Hypothesis 3) *Impatiens glandulifera* has a relatively more within population variation.

The average number of alleles per locus per population was greater 4.70 for *I. glandulifera* and 3.98 for *H. mantegazzianum* (Sections 4.1 and 5.1). Therefore this hypothesis was accepted, as *I. glandulifera* has proportionally more variation within a population. This result was consistent with the finding that there was more temporal variation in populations of *I. glandulifera* and reflects the larger population sizes (so each sampled population was a smaller proportion of the number of individuals than with *H. mantegazzianum*).

An unexpected finding of this study was the high amount of within population variation found in *H. mantegazzianum.* Many populations were composed of 40 or fewer individuals and the species is self-compatible and capable of producing thousands of seeds per individual. Therefore, populations could have arisen from just one founder. However, this study found all populations to have at least five alleles at locus A34, so this was not found to be the case (Figure 4.2).

Hypothesis 4) *Impatiens glandulifera* has a greater proportion of overall genetic variation due to between catchment (rather than within catchment) variation than *H. mantegazzianum.*

This hypothesis was proposed because *I. glandulifera* has a greater dispersal ability than *H. mantegazzianum* and its pollinators have larger home ranges (within catchments). Therefore, populations of *I. glandulifera* in the same catchment would be expected to be more similar than those of *H. mantegazzianum* and a higher proportion of the total variation in *I. glandulifera* would be due to between, rather than within catchment variation. However, the F_{ST} values comparing catchments were not greater for *I. glandulifera* than for *H. mantegazzianum* and assignment tests between catchments found there to be clearer differences in the case of *H. mantegazzianum.* Therefore, this hypothesis is rejected.

There are a number of possible reasons for this finding. *Impatiens glandulifera* is a much more popular garden plant and so it is likely to have been introduced into each catchment on many more occasions than *H. mantegazzianum*, which would lead to there being a greater amount of within catchment variation in *I. glandulifera.* There are many more individuals of *I. glandulifera* present in each catchment (as population sizes are much greater and the overall distribution of the species is much wider) and there would be expected to be a positive correlation between number of and amount of variation. In addition, *I. glandulifera* is an annual and so has a much higher turnover rate than *H. mantegazzianum*. Therefore, a higher level of gene flow may be required to prevent population differentiation between two populations as the effects of drift may take hold much faster.

An additional point of interest revealed by this study was the low level of cpDNA variation observed. Only one locus was found to be polymorphic for each species (out of three microsatellite loci and two other variable regions tested (section 3.3)) and only two alleles were found in *I. glandulifera.* All populations of *H. mantegazzianum* were monomorphic for the same allele, except population Tyne Ouse 4, which had three private alleles and did not contain the allele present in all other populations. This population appears to have been introduced from a source or sources that were different from that of all other populations (including that from London). The lack of variation could be a reflection of number of individuals and native range from which introduced individuals arose.

The identification of the number and location of introductions in the Northeast of England would enable models of the spread of the species to be more accurate. Populations that appear to have arisen from independent introductions were identified in both species. The populations of *H. mantegazzianum* were Tees 59, Tees 162, Wear

77, Wear 18, Tyne 94 and Tyne Ouse 4. The populations of *I. glandulifera* were Tees 189, Wear Rainton 52, Tyne 92 and Tyne Ouse 4.

6.10 Further Work

A large amount of genetic variation was observed in both species, and this suggests that either there have been numerous introductions of the species into the UK or that large numbers of individuals were present in initial introductions from their native source. In order to put the amount of variation in populations in the study area into perspective, a comparison with populations in native areas could help assess the number of source populations. Comparisons of heterozygosity levels in native populations with those observed in this study would enable comparisons of levels of selfing to be compared as increased inbreeding has been found in many introduced populations of other species.

More extensive screening of potential sources of introduction from within the UK would also be useful. It would show how much variation there was within the UK as a whole and may also enable the identification of the sources of some of the populations screened.

An investigation of the behaviour and home range of pollinators of the species could provide information on the levels of and distance of gene flow between populations. This may also be useful in determining levels of inbreeding to see if they match apparent levels of inbreeding observed by deviations of heterozygosity levels of microsatellite alleles from the Hardy-Weinberg equilibrium. Many species of insects are thought to pollinate *H. mantegazzianum*, but the relative proportions of the species involved are not known, and could be determined by observational studies.

Both species appear to have been introduced into the Northeast of England on many occasions and from a number of different sources. There is a lack of information on the source of the introduced material and on introductions. An investigation of historical records may provide useful information as to the source of introductions. The first records of introductions of both species into the UK were in the south of the country but it is possible that the species was brought into the Northeast directly from native areas. For example, there may have been introductions from ships arriving in any of the three catchments.

When considering the causes of variation between populations, it was often difficult to distinguish between variation due to independent introductions and that due to drift where gene flow between populations was low. This problem may be aided by sampling a greater number of individuals at each population, and by sampling populations in between those already genotyped.

References

Akkaya, M.S, Bhagway, A.A. and Cregan, P.B. (1992) Length polymorphism of simple sequence repeat in DNA in soybean. *Genetics,* **132**, 1131-1139.

Alexandrino, J., Froufe, E., Arntzen, J.W. and Ferrand, N. (2000) Genetic subdivision, glacial refugia and postglacial recolonization in the golden-striped salamander, *Chioglossa lusitanica* (Amphibia: Urodela). *Molecular Ecology,* **9**, 771-781.

Amsellem, L, Noyer, J.L., Le Bourgeois, T. and Hossaert-McKey, M. (2000) Comparison of genetic diversity of the invasive weed *Rubus alceifolius* Poir. (Rosaceae) in its native range and in areas of introduction, using amplified fragment length polymorphism (AFLP) markers. *Molecular Ecology,* **9**, 443-455.

Aquadro, C.F., Noon, W.A. and Begun, D.J. (1992) RFLP analysis using heterologous probes. *Molecular genetic analysis of populations. A practical approach* (ed A.R. Hoelzel) pp. 115-157. Oxford University press.

Arora, K., Grace, C. and Stewart, F (1982) Epidermal features of *Heracleum mantegazzianum* Somm. and Lev., *H. sphondylium* L. and their hybrid. *Botanical Journal of the Linnean Society,* **85**, 169-177.

Beaumont, M.A. (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013-2029.

Beerling, D.J. (1993) The impact of temperature on the northern distribution limits of the introduced species *Fallopia japonica* and *Impatiens glandulifera* in north-west Europe. *Journal of Biogeography*, **20**, 45-53.

Beerling, D.J. and Perrins, J.M. (1993) *Impatiens glandulifera* Royle (*Impatiens roylei* Walp.). *Journal of Ecology*, **81**, 367-382.

Birky, C.W., Fuerst, P. and Maruyama, T. 1989 Organelle gene diversity under migration, mutation and drift equilibrium expectations, appreach to equilibrium ,

effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics*, **121**, 613-627.

Birnboim, H.C. and Doly (1979) A rapid alkaline exraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research*, **7**, 1513-1523.

Bjorkland, M.J. and Baker, A.J. (1996) The successful founder: Genetics of introduced *Carduelis chloris* (greenfinch) populations in New Zealand. *Heredity,* **77**, 410-422.

Bodmer, W.F. and Cavalli-Sforza (1967) A migration matrix model for the study of random genetic drift. *Genetics*, **59**, 565-592.

Boecklen, W.J. and Howard, D.J. (1997) Genetic analysis of hybrid zones: number of markers and power of resolution. *Ecology*, **78,** 2611-2616.

Bossart, J.L and Prowell, D.P (1998) Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology and Evolution,* **13**, 202-206.

Callen, D.F., Thompson, A.D., Shen, Y., Phillips, H.A., Richards, R.I., Mulley, J.C., and Sutherland, G.R. (1993) Incidence and origin of 'null' alleles in the (AC)n microsatellite marker. *American Journal of Human Genetics*, **52**, 922-927.

Caron, H., Dumas, S., Marque, G., Messier, C., Bandou, E., Petit, R.J., and Kremer, A. (2000) Spatial and temporal distribution of chloroplast DNA polymorphism in a tropical tree species. *Molecular Ecology,* **9**,1089-1098.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic analysis models and estimation procedures. *American Journal of Human Genetics,* **19**, 233-257.

Chase, M., Kessel, R. and Bawa, K. (1996) Microsatellite markers for population and conservation genetics of tropical trees. *American Journal of Botany*, **83**, 51-57.

Chase, M.R, Boshier, D.H. and Bawa, K.S. (1995) Population genetics of *Cordia alliodora* (Boraginaceae), a neotropical tree. 1 Genetic variation in natural populations. *American Journal of Botany,* **82**, 468-475.

Chee, P., Kapaum, J. and Zamora-Diaz, M. (1996) Microsatellite DNA as a genetic marker. *North Dakota State University, website*: www.ndsu.nodak.edu/instruct/mcclean/plsc731/vntr.htm

Chung, C.T. and Miller, R.H. (1988) A rapid and convenient method for the preparation and storage of competent bacterial cells. *Nucleic Acids Research,* **16**, 3580.

Cockerham, C.C. and Weir, B.S. (1993) Estimation of gene flow from *F*-statistics. *Evolution*, **47**, 855-863.

Collingham, Y.C., Huntley, B. and Hulme, P.E. (1997) The use of a spatially explicit model to simulate the spread of a riparian weed. *Species dispersal and Land Use Processes* (eds A. Cooper and J. Power), pp. 45-52. Proceedings of the International Association for Landscape Ecology, IALE (UK), Belfast.

Collingham, Y.C., Wadsworth, R.A., Huntley, B. and Hulme, P.E. (2000) Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology,* **37**, 13-27.

Comes, H.P. and Abbott, R.J. (1998) The relative importance of historical events and gene flow on the population structure of a Mediterranean ragwort, *Senecio gallicus* (Asteraceae). *Evolution,* **52**, 355-367.

Condit, R. and Hubbell, S.P. (1991) Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. *Genome,* **34**, 66-71.

Crawley, M.J. (1986) The population biology of invaders. *Philosophical Transactions of the Royal Society of London Series B*, **314**, 711-731.

Crawley, M.J. (1989a) Invaders. *Plants Today,* **October**, 152-158.

Crawley, M.J. (1989b) Chance and timing in biological invasions. *Biological Invasions: a Global Perspective* (eds Drake, JA. *et al.*), pp 407-423. New York, John Wiley and Sons Ltd.

Davies, N. Villablanca, F.X. and Roderick, G.K. (1999) Determining the source of inidividuals: multilocus genotyping in non equilibrium population genetics. *Trends in Ecology and Evolution*, **14**, 17-21.

Dawson, F.H. and Holland, D. (1999) The distribution in bankside habitats of three alien invasive plants in the U.K. in relation to the development of control strategies. *Hydrobiologia,* **415,** 193-201.

Demesure, B., Sodzi, N. and Petit, R.J. (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Molecular Ecology*, **4**, 129-131.

Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M., and Freimer,N.B. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Science of the USA,* **91**, 3166-3170.

Drummond, R.S.M., Keeling, D.J., Richardson, T.E., Gardner, R.C. and S.D. Wright (2000) Genetiv analysis and conservation of 31 surviving individuals of a rare New Zealand tree, *Metrosideros bartlettii* (Myrtaceae). *Molecular Ecology,* **9**, 1149-1157.

Dunn, T.C. (1977) Notes: pollination of Himalyan Balsam. *The Vasculum,* **62,** 61.

Dutech, C., Maggia, L., and Joly, H.I. (2000) Chloroplast diversity in Vouaccapoua americana (Caesalpiniaceae), a neotropical forest tree. *Molecular Ecology,* **9**, 1427-1432.

Edwards, A., Hammond, H.A., Jin, L., Caskey, C.T., and Chakraborty, R. (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics,* **12**, 241-253.

Eguiarte, L.E., Perez-Nashier, N. and Piñero, D. (1992) Genetic structure, outcrossing rate and heterosis in *Astrocaryum mexicanum* (tropical palm): implications for evolution and conservation. *Heredity*, **69**, 217-229.

Ennos, R.A. (1994) Estimating the relative rates of pollen and seed migration among plant populations. *Heredity*, **72**, 250-259.

Felsenstein, J. (1981) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution,* **35**, 1229-1242.

Felsenstein, J. (1984) Distance methods for inferring phylogenies: a justification. *Evolution,* **38**, 10-24.

Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Fisher, D. and Bachmann, K (1998) Microsatellite enrichment in organisms with large genomes (*Allium* cepa L). *Biotechniques,* **24** 796-799.

Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science,* **155**, 279-284.

Freckleton, R.P. and Watkinson, A.R. (1998) How does temporal variability affect predictions of weed population numbers? *Journal of Applied Ecology*, **35**, 340-344.

Fu, Y-H., Kuhl, D.P.A., Pizzutii, M. Pieretti, M., Sutcliffe, J.M., Richards, S., Verkerk, A., Holden, J., Fenwick, R., and Warren, S.T. (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell,* **67**, 1047-1058.

Gaggiotti, O.E., Lange, O., Rassmann, K and Gliddon, C. (1999) A comparison f two indirect methods for estimating average levels of gene flow using microsatellite data. *Molecular Ecology* **8**, 1513-1520.

Giles, B.E. and Goudet, J. (1997) Genetic differentiation in *Silene dioica* metapopulations: estimation of spatiotemporal effects in a successional plant species. *American Naturalist*, **149**, 507-526.

Goldstein, D.B., Linares, A.R., Cavalli-Sforza, L.L., and Feldman, M.W. (1995a) An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**, 463-471.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. (1995b) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the USA*, **88**, 335-342.

Goodman, S.J. (1997) RST CALC: A collection of computer programs for calculating unbiased estimates of genetic differentiation and determining their significance for microsatellite data. *Molecular Ecology*, **6**, 881-885.

Govindaraju, D.R. (1988) Relationship between dispersal ability and levels of gene flow in plants. *Oikos*, **52**, 31-35.

Grace, J. and Nelson, M. (1981) Insects and their pollen loads at a hybrid *Heracleum* site. *New Phytologist,* **87**, 413-423.

Grime, J.P., Hodgson, J.G. and Hunt, R. (1988) Comparitive plant ecology: A functinal approach to common British species. London, Unwin Hyman.

Guo, W.S. and Thompson, E.A. (1992) Performing the exact test of Hardy-Weinberg Proportion for Multiple Alleles. *Biometrics,* **48**, 361-372.

Hamrick, J.L. (1982) Plant Populations Genetics and Evolution. *American Journal of Botany,* **69**, 1685-1693.

Hamrick, J.L. and Godt M.J.W. (1996) Endemic Plants. *Conservation genetics- case histories from nature* (ed. J.C. Avis) pp 281-304. London, Chapman and Hall.

Higgins, S.I., Richardson, D.M. and Cowling, R.M. (1996) Modeling invasive plant spread: the role of plant-environment interactions and model structure. *Ecology,* **77**, 2043-2054.

Higgins, S.I. and Richardson, D.M. (1999) Predicting plant migration rates in s changing world: the role of long-distance dispersal. *American Naturalist,* **153**, 464-475.

Holdgate, M.W. (1986) Summary and conclusions: characteristics and consequences of biological invasions. *Philosophical Transactions of the Royal Society of London Series B*, **314**, 733-742.

Hood, W.G. and Naiman, R.J. (2000) Vulnerability of riparian zones to invasion by exotic vascular plants. *Plant Ecology* **148**, 105-114.

Hu, X.S. and Ennos, R.A. (1997) On estimation of the ratio of pollen to seed flow among plant populations. *Heredity* **79**, 541-552.

Hulme, P.E., Huntley, B., Wyatt, B., Preston, C., Hoelzel, A.R., Blundel, C. Collingham, Y., Wadsworth, R. and Willis, S.G. (1999) Spatial and temporal scale dependencies in the invasion of riparian habitats by alien weeds and validating models of species spread in riparian habitats using molecular markers. *NERC Large Scale Processes in Ecology and Hydrology final report.*

Ibrahim, K.M., Nichols, R.A. and Hewitt, G.M. (1996) Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity,* **77**, 282-291.

Jarne, P and Lagoda, J.L. (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution,* **11**, 424-429.

Keys, R.N. and Smith, S.E. (1994) Mating system parameters and population genetic structure in pioneer populations of *Prosopis velutina* (Leguminosae) *American Journal of Botany,* **81**, 1013-1020.

Kimura, M. and Crow, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics,* **49**, 725-738.

Kirkpatrick, J. (1994) A continent transformed. Human impact on the natural vegetation of Australia p 92-95.. Australia, Oxford University Press.

Koblížková, A., Doležel, J. and Macas, J. (1998) Subtraction with 3′ modified oligonucleotides eliminates amplification artefacts in DNA libraries enriched for microsatellites. *BioTechniques,* **25**, 32-38.

Kvist, L., Ruokonen, M., Lumme, J. and Orell, M. (1999) The colonization history and present-day population structure of the European great tit (*Parus major major*). *Heredity,* **82**, 495-502.

Lagercrantz, U., Ellegren, H. and Andersson, L. (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucleic Acids Research,* **21**, 1111-1115.

Lammi, A., Siikamaki, P. and Mustajarvi, K. (1999) Genetic diversity, population size, and fitness in central and peripheral populations of a rare plant *Lychnis viscaria. Conservation Biology*, **13,** 1069-1078.

Lehmann, T., Hawley, W.A., Kamau, L., Fontenilles, D., Simards, F. and Collins, F.H. (1996) Genetic differentiation of *Anopheles gambiae* populations from East and West Africa: comparison of microsatellite and allozyme loci. *Heredity*, **77**, 192-208.

Levinson, G. and Gutman, G.A. (1987) Slipped-strand misrepairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, **4**, 203-221.

Lovei, G.L. (1997) Global change through invasion. *Nature*, **388**, 627.

Luikart, G., Allendorf, F.W., Cornuet, J.-M. and Sherwin, W.B. (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. *The Journal of Heredity*, **89**, 238-247.

Mack, R.N. (1985) Invading plants: their potential contribution to population biology. *Studies on plant demography* (eds Festschrift, A. and J.L. Harper), pp.127-142. London, Academic press Inc. (London) Ltd.

Mahy, G., Ennos, R.A. and Jacquemart, A.L. (1999) Allozyme variation and genetic structure of *Calluna vulgaris* (heather) populations in Scotland: the effect of postglacial recolonization. *Heredity,* **82**, 654-660.

Maki, M., Morita, H., Oiki, S. and Takahashi, H. (1999) The effect of geographic range and dichogamy on genetic variability and population genetic structure in *Tricyrtis* section *Flavae* (Liliaceae). *American Journal of Botany*, **86,** 287-292.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.

Martínez-Palacios, A., Eguiarte, L.E. and Furnier, G.R. (1999) Genetic diversity of the endangered endemic *Agave victoriae-reginae* (Avagaceae) in the Chihuahuan desert. *American Journal of Botany*, **86**, 1093-1098.

Mason-Gamer, R.J., Holsinger, K.E. and Jansen, R.K. (1995) Chloroplast DNA haplotype variation within and among populations of *Coreopsis glandiflora* (Asteraceae). *Molecular Biology and Evolution,* **12**, 371-385.

McCauley, D.E. (1995) The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends in Ecology and Evolution*, **10**, 198-202.

McCauley, D.E., Raveill, J. and Antonovics, J. (1995) Local founding events as determinants of genetic structure in a plant metapopulation. *Heredity*, **75**, 630-636.

McClintock, D. (1975) *Heracleum.* In *Hybridisation and the Flora of the British Isles.* C.A. Stace (Ed.) pp.270. London, Academic Press.

Michalalkis, Y. and Excoffier, L. (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics,* **142**, 1061-1064.

Milligan, B.G. (1997) Total DNA isolation. *Molecular genetic analysis of populations. A practical approach* (ed. Hoelzel, A.R.). Second edition, Oxford, Oxford University press pp.445.

Milne, R.I. and Abbott, R.J. (2000) Origin and evolution of invasive naturalized material of *Rhododendron ponticum* L. in the British Isles. *Molecular Ecology,* **9**, 541-556.

Mogensen, H.L. (1996) The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany,* **83**, 383-404.

Moody, M.E. and Mack, R.N. (1988) Controlling the spread of plant invasions: the importance of nascent foci. *Journal of Applied Ecology*, **25**, 1009-1021.

Moore, S.S., Sargeant, L.L., King, T.J., Mattick, J.S., Georges, M. and Hetzel, D.J.S. (1991) The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. *Genomics*, **10**, 654-660.

Murray, M.G. and Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, **8**, 4321-4325.

Nathan, R. and Muller-Landau, H.C. (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution* **15**, 278-282.

Nei, M. (1972) Genetic distance between populations. *The American Naturalist,* **106**, 283-292.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Science of the USA,* **70,** 3321-3323.

Neuffer, B and Hurka, H. (1999) Colonization history and introduction dynamics of *Capsella bursa-pastoris* (Brassicaceae) in North America: isozymes and quantitative traits. *Molecular Ecology,* **8**, 1667-1681.

Nickrent. D.L. (1994) From field to film: rapid sequencing methods for field-collected plant species. *BioTechniques*, **16**, 470-472.

Novak, S.J. and Mack, R.N. (1993) Genetic variation in *Bromus tectorum* (Poaceae): comparison between native and introduced populations. *Heredity,* **71**, 167-176.

Novak, S.J. and Mack, R.N. (1995) Allozyme diversity in the apomictic vine *Bryonia alba* (Cucurbitaceae)- potential consequences of multiple introductions. *American Journal of Botany,* **82**, 1153-1162.

Novak, S.J. and Welfley, A.Y. (1997) Genetic diversity in the introduced clonal grass *Poa bulbosa* (Bulbous bluegrass). *Northwest Science,* **71**, 271-280.

Paetkau, D (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology,* **4**, 347-354.

Pemberton, J.M., Slate, J., Bancroft, D.R., and Barrett, J.A. (1995) Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. *Molecular Ecology,* **4,** 249-252.

Perez de la Vega, M., Garcia, P. and Allard, R.W. (1991) Multilocus genetic structure of anvestral Spanish and colonial Californian populations of *Avena barbata. Proceedings of the National Academy of Science of the USA,* **88,** 1202-1206.

PerezLezaun A, Calafell F., Mateu E., Comas D., RuizPacheco R., and Bertranpetit J. (1997) Microsatellite variation and the differentiation of modern humans. *Human Genetics*, **99**, 1-7.

Perrins, J., Fitter, A. and Williamson, M. (1993) Population biology and rates of invasion of three introduced *Impatiens* species in the British Isles. *Journal of Biogeography*, **20**, 33-44.

Pritchard, J.K. and Feldman, M.W. (1996) Statistics for microsatellite variation based on coalescence. *Theoretical Population Biology*, **50**, 325-344.

Pysek, P. (1991) *Heracleum mantegazzianum* in the Czech Republic- dynamics of spreading form the historical perspective. *Folia Geobotanica and Phytotaxonomica*, **26**, 439-454.

Pysek, P and Prach, P (1993) Plant invasions and the role of riparian habitats: a comparison of four species alien to central Europe. *Journal of Biogeography*, **20**, 413-420.

Pysek, P and Pysek, A. (1995) Invasion by *Heracleum mantegazzianum* in different habitats in the Czech Republic. *Journal of Vegetation Science*, **6**, 711-719.

Queller, D.C., Strassmann, J.E. and Hughes, C.R. (1993) Microsatellites and kinship. *Trends in Ecology and Evolution*, **8**, 385-388.

Raspé, O., Saumitou-Laprade, P., Cuguen and Jacquemart, A.-L. (2000) Chloroplast DNA haplotype variation and population differentiation in *Sorbus aucuparia* L. (Rosaceae: Maloideae). *Molecular Ecology*, **9**, 1113-1122.

Raybould, A.F., Mogg, R.J., Aldam, C., Gliddon, C.J., Thorpe, R.S. and Clarke, R.T. (1998) The genetic structure of sea beet (*Beta vulgaris ssp. maritima*) populations. III. Detection of isolation by distance at microsatellite loci. *Heredity*, **80**, 127-132.

Raybould, A.F., Mogg, R.J., Clarke, R.T., Gliddon, C.J. and Gray, A.J. (1999) Variation and population structure at microsatellite and isozyme loci in wild cabbage (*Brassica oleracea* L.) in Dorset (UK). *Genetic Resources and Crop Evolution* **46**, 351-360.

Raymond M. and Rousset F, (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Reusch, T.B.H., Stam, W.T. and Olsen, J.L. (2000) A microsatellite-based estimation of clonal diversity and subdivision in *Zostera marina*, a marine flowering plant. *Molecular Ecology*, **9**, 127-140.

Rice, W.R. (1989) Analysing tables of statistical tests. *Evolution*, **43**, 223-225.

Richards, C.M., Church, S. and McCauley, D.E. (1999) The influence of population size and isolation on gene flow by pollen in *Silene alba*. *Evolution*, **53**, 63-73.

Robertson, A. and Hill, W.G. (1984) Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics*, **107**, 703-718.

Rousset F, 1996. Equilibrium values of measure of population subdivision for stepwise mutation processes. *Genetics*, **142**, 1357-1362.

Rousset, F. (1997) Genetic differentiation and estimation fo gene flow from *F*-statistics under isolation by distance. *Genetics*, **145**, 1219-1228.

Sahai Maroof, M.A., Biyashev, R.M., Yang, G.P., Zhang, Q. and Allard, R.W. (1994) Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. *Proceedings of the National Academy of Sciences USA*, **91**, 5466-5470.

Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* **4**, 406-425.

Sambrook, J., Fritsch, E.F. and Maniatis, T (1989) *Molecular cloning. A laboratory manual (2nd edn.).* Cold Spring Harbor, New York, Cold Spring Harbor Laboratory press.

Schlötterer, C. (1997) Total DNA isolation. *Molecular genetic analysis of populations. A practical approach* (ed Hoelzel, A.R.). Second edition, Oxford, Oxford University press pp.445.

Schneider, S., Roessli, D., and Excoffier, L (2000*) Arlequin ver. 2.000: A software for population genetics data analysis.* Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Shriver, M.D., Jin, L., Chakraborty, R., and Boerwinkle, E. (1993) VNTR allele frequency distributions under the SMM: A computer simulation approach. *Genetics,* **134**, 983-993.

Silvertown, J. (1991) Dorothy's dilemma and the unification of plant population biology. *Trends in Ecology and Evolution*, **6**, 346-348.

Slatkin, M. (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264-279.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457-462.

Slatkin, M. and Barton, N.H. (1989) A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, **43**, 1349-1368.

Skogland, S.J. (1990) Seed dispersing agents in two regularly flooded river sites. *Canadian Journal of Botany*, **68**, 754-760.

Soltis, D.E., Soltis, P.S. and Milligan, B.G. (1992) Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. *Molecular systematics of plants* (eds Soltis, P.S., Soltis, D.E. and J.J. Doyle), pp.117-150. London, Chapman and Hall.

Stace, C. (1991) *New Flora of the British Isles.* Cambridge, Cambridge University Press.

Strand, A.E, Milligan, B. G. and Pruitt, C.M. (1996) Are populations islands? Analysis of chloroplast DNA variation in *Aquilegia. Evolution,* **50**, 1822-1829.

Sun, M. (1997) Genetic diversity in three colonizing orchids with contrasting mating systems. *American Journal of Botany,* **84**, 224-232.

Sweigart, A., Karoly, K., Jones, A. and Willis, J.H. (1999) The distribution of individual inbreeding coefficients and pairwise relatedness in a population of *Mimulus guttatus. Heredity,* **83**, 625-632.

Taberlet, P., Gielly, L., Pautou, G. and Bouvet, J. (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, **17**, 1105-1109.

Tiley, G.E.D., Dodd, F.S. and Wade, P.M. (1996) *Heracleum mantegazzianum* Sommier and Levier. *Journal of Ecology*, **84**, 297-319.

Trewick, S. and Wade, P.M. (1986) The distribution and dispersal of two alien species of *Impatiens*, waterway weeds in the British Isles. *Proceedings EWRS/AAB 7th Symposium of aquatic weeds*, 351-356.

Tufto, J. Engen, S. and Hindar, K. (1996) Inferrring patterns of migration from gene frequencies under equilibrium conditions. *Genetics*, **144**, 1911-1921.

Turner, M.E., Stephens, J.C. and Anderson, W.W. (1982) Homozygosity and patch structure in plant populations as a result of nearest-neighbor pollination. *Proceedings of the National Academy of Sciences of the USA,* **79**, 203-207.

Valdes, A.M., Slatkin, M., and Freimer, N.B. (1993) Allele frequencies at microsatellite loci: the SMM revisited. *Genetics,* **133,** 737-749.

Van der Merwe, M., Winfield, M.O., Arnold, G.M. and Parker, J.S. (2000) Spatial and temporal aspects of the genetic structure of *Juniperus communis* populations. *Molecular Ecology* **9**, 379-386.

Viard, F, Justy, F. and Jarne, P. (1997) The influence of self-fertilization and population dynamics on the genetic structure of subdivided populations: a case study using microsatellite markers in the freshwater snail *Bulinus truncatus*. *Evolution*, **51,** 1518-1528.

Vogt Anderson, V. and Calor, B. (1996) Long term effects of sheep grazing on giant hogweed (*Heracleum mantegazzianum*). *Hydrobiologia,* **340**, 277-284.

Wade, M.J. and McCauley, D.E. (1988) Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution*, **42**, 995-1005.

Wadsworth, R.A., Swetnam, R.D. and Willis, S.G. (1997) Seeds and sediment: modelling the spread of Impatiens glandulifera Royle. *Proceedings of the sixth annual IALE(UK) conference*, **Ulster**, 53-60.

Wadsworth, R.A., Collingham, Y.C., Willis, S.G., Huntley, B. and Hulme, P.E. (2000) Simulating the spread and management of alien riparian weeds: are they out of control? *Journal of Applied Ecology,* **37**, 28-38.

Wang, J. (1997) Effective size and F-Statistics of subdivided populations. I Monoecious species with partial selfing. *Genetics,* **146**, 1453-1463.

Weber, J.L. and Wong, C. (1993) Mutation of human short tandem repeats. *Human. Molecular Genetics,* **2**, 1123-1128.

Weimark, G., Stewart, F., and Grace, J. (1979) Morphometric and chromatographic variation and male meiosis in the hybrid *Heracleum mantegazzianum* × *H. sphondylium* (Apiaceae) and its parents. *Hereditas,* **91**, 117-127.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution,* **38**, 1358-1370.

Weising, K. and Gardner, R.C. (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome*, **42**, 9-19.

Whitlock, M.C. (1992) Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution,* **46**, 608-615.

Whitlock, M.C. and McCauley, D.E. (1990) Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution*, **44**, 1717-1724.

Whitlock, M.C. and McCauley, D.E. (1999) Indirect measures of gene flow and migration $F_{ST} \neq 1/(4Nm + 1)$. *Heredity,* **82**, 117-125.

Williamson, M.H. (1998) Biological Invasions. London, Chapman and Hall. 256pp.

Williamson, M.H. and Brown, K.C.(1986) The analysis and modelling if British invasions. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences,* **314**, 505-512.

Willis, S.G. (1999) Factors affecting the distribution of three non-indigenous riparian weeds in North-East England. Ph.D. thesis, University of Durham.

Wong, K.C. and Sun, M. (1999) Reproductive biology and conservation genetics of *Goodyera procera* (Orchidaceae). *American Journal of Botany,* **86**, 1406-1413.

Wright, S. (1943) Isolation by distance. *Genetics*, **28**, 114-138.

Wright, S. (1978) *Evolution and the Genetics of Populations.* **Vol. 4**. University of Chicago Press, Chicago Ill., USA.

Young, A.G. and Murray, B.G. (2000) Genetic bottlenecks and dysgenic gene flow into re-established populations of the grassland daisy, *Rutidosis leptorrhynhoides. Australian Journal of Botany,* **43**, 409-416.